How Well Do We Know It?

S.T.LAKSHMIKUMAR

How Well Do We *Know* It?

How Well Do We Know It?

S T Lakshmikumar

PUBLICATION

Published by The Author
Cover Design : E. Arul

stlakshmikumar@gmail.com

Contents

नमो गुरुभ्यः

Acknowledgements

I am deeply indebted to
everyone who helped me learn
through
teaching, discussion and argument.

This book is the culmination of my effort as a scientist to understand the world at large. The inspiration came from various experiences and comments of Richard Feynman as recorded in his memoirs, nonprofessional writings, stories and anecdotes. With this guiding spirit, distillation of the knowledge acquired from my experiences and reading into a coherent philosophy was possible. This is an attempt at a grand synthesis, made by an individual with no accomplishments to serve as a halo of wisdom. I have little confidence that I have succeeded and much less that these views would be palatable to most.

I do not claim that all ideas expressed here are original. But I cannot provide a complete list of sources that have influenced my presentation, since I never maintained any such record. Wikipedia served as a great source of basic information in many of the areas discussed here. The data of carbon dioxide concentration and temperature anomaly presented in chapter XII as also the data sets of Anscombe discussed in chapter V are extensively and freely available in web pages. I am grateful to the authors for the free availability of these data. I have tried my level best not to use sentences of others verbatim except when I attributed them to the author. The sources for these quotations may not have been included in the short bibliography that has been provided at the end. It is as possible that I have misquoted

as to have not given proper attribution. In both cases I can only offer my apologies and promises to correct them in case there are future editions.

The proximate cause for whatever was accomplished in this effort are the extended series of long interactions I had with Dr. S. Mohan. As I continue to regret the brief tenure of about three years during which this close association lasted, I am more indebted to that experience than can be expressed in these few words of acknowledgement. I am also grateful for his effort as a copy editor. I am also indebted to Dr. S. M. Shivaprasad who over many years served as a sounding board for many of my ambitious forays into grand interpretations.

I am deeply indebted to Dr. Amitabha Basu for a painstaking reading of the draft and numerous suggestions and comments. I would like to thank Dr. T. D. Senguttuvan for help during the later stages of preparing the manuscript and to Mr. E. Arul for designing the book cover. I am obliged to old friend Dr. A. V. Subbarao for critical comments on medical issues discussed herein and to Dr. P. N. Vijay Kumar, Dr. R. S. Arora and other colleagues at NPL for discussions, suggestions and comments. I am grateful to the National Physical Laboratory and to the successive directors of NPL for enabling my growth as a scientist all these years.

S T Lakshmikumar

November 2011
New Delhi
lakshmikumar@nplindia.org
stlakshmikumar@gmail.com

1

HOW FAR DOES THIS DISCUSSION GO?

1.1 What is being attempted

"How well do we know it"? The question assumes that there is something to be known and that something can be known. Precisely describing "it" is the first step of knowing it. However, precise definitions using words are not enough, not at least for the present discussion. Richard Feynman describes in his famous memoirs, "Surely you are joking Mr. Feynman", his hilarious experience with a group of philosophy students. Philosophers try to define things precisely and endlessly polish the statements and definitions. In this case they were discussing Whitehead's theory of essential objects and the book "Process and Reality". The problem with definitions became apparent when the elaborate description did not help the students to decide if a "brick" was an essential object. Not withstanding his disparaging approach to philosophy, when he tried his hand at Chinese calligraphy and art, Feynman found he was able to say "it is good" or "it is not good" without being able to explain "why". Thus we can "know" things even if we cannot define them precisely. The problem was known to philosophers since antiquity and it is not to be expected that this small monograph will resolve such issues.

In the interest of saying something sensible about "how well we know", "it" is confined to subjects with objective descriptions. At every stage the description is justified with examples. Examples do not guarantee that the descriptions are "objective". But a large number of examples will hopefully help in accepting that the present analysis is useful. Consensual or nearly universal acceptance will be possible. It is not possible to realize "how well" we know "it" in the absence of this restriction.

In the restricted domain being considered, quantification is accepted as the key to improving "how well" we know. We will consider the use of numbers and mathematics to draw conclusions about how well we know. Hopefully, these will also be accepted by most. After all it is quite possible to find a person who adamantly refuses to accept that 2 added to 2 makes 4. So no conclusion will be universally acceptable.

This does not mean that things which cannot be quantified do not exist or even that they are not important. Merely that they will not be discussed here. The core objective is to investigate the use and misuse of quantification during the process of improving how well we know. The goal is to be able to evaluate relative strengths of two statements and be confident that the perceived superiority of one over the other is real.

What is being attempted is best illustrated by the following statements. (i) The sun will rise tomorrow. (ii) It is predicted beforehand that the sun will rise on 2nd September 2009 at 6:22 AM IST in Surat, 6:00AM IST in Delhi and 5:16 AM IST in Darjeeling. (iii) It is predicted beforehand that the sun will rise on 22nd July 2009 at 6:08 AM IST in Surat, 5:36 AM IST in Delhi 4:55 AM IST in Darjeeling. However, sun will not be visible at Surat during 6:22:54-6:24:33 and at Darjeeling during 6:27:01-6:28:30 as there will be a total solar eclipse. The eclipse at Delhi will be only partial and not total. Obviously, in these statements, there is progressive improvement in how well we know about the events taking place at sunrise. Near universal consensus of this progressive improvement can be intuitively accepted. The discussion in the following pages is applicable only to statements of this nature. Quantification using numbers and

mathematics is the key to this improvement. The goal is to precisely examine the various ways in which this is accomplished and understand the strengths and weaknesses of this approach.

A few issues have been ignored. Consider the first statement. That the sun will rise tomorrow is generally believed to be true since this has been observed everyday in the past. However, a philosopher questions this process. There is no logical reason why the fact that sunrise was observed in the past should ensure that it will do so tomorrow. They call this an "induction" which is not logically justified. We will look at the problem in the context of the basic theories of physics later. However, the goal of the present discussion is practical rather than philosophical. The philosophical discussion of the problem of induction is ignored. But comparison of practical statements will be considered even if individually the statements are conclusions drawn from induction. There will be no attempt to discuss why most people given the earlier example will agree that there is a progressive improvement. This is a relevant but independent scientific question of human psychology and communication.

The second of the above statements could be made if high quality clocks are available at all locations and there is enough confidence that they all show the same time. The experimental issues in this are not of present concern. How these instruments are built and how one creates confidence in their performance is not relevant. The discussions in these pages also ignore whether the data is collected by human perception or by automatic equipment. Thus, limits due to human capabilities of perception will not be discussed.

Determining the time of sunrise can be done objectively. It can be ensured that most people would agree with the observations. To predict the time of sunrise beforehand however, gradual variation of the time of sunrise over the years has to be monitored and analyzed. There can be various methods to accomplish this. The relative capabilities of these methods are the subject of the present enquiry.

Predicting solar eclipses would require understanding the movement of the earth, sun and moon. It is not however necessary to accept the modern sun centered planetary system. For example, ancient

Indian astronomy based on simple extension of the geocentric system of Ptolemy provides a means of predicting the timing of a solar eclipse with an accuracy of about 30 min. Complex mathematical physics using Kepler's laws and Newtonian mechanics enhances the accuracy of the predictions. How far observations justify the use of more and more complex mathematics is the topic under discussion.

Consider as another example the observations on planet Venus by the ancient Mayan civilization. They observed a morning star for 236 days, then it cannot be seen for 90 days. Then they observed an evening star for 250 days and finally it disappears again for 8 days. This would have enabled them to predict in advance when there would be an evening star or when there would be none. They could have determined the total period of the appearance of the stars as 584 days. By repeated measurements over centuries they probably found that the cycle is not exactly 584 days long but a fractional value close to it. They may have guessed that both the evening and morning stars are the same, namely planet Venus.

If we use Kepler's laws of planetary motion, we understand more. Venus is not observed when it is between the earth and the sun or on the far side of the sun. The number of days of absence is different because Venus appears to move more slowly in the sky when it is on the far side of the sun compared to when it passes between the earth and the sun. One can not only predict the presence or absence but also the position in the sky. Obviously a more sophisticated mathematical description such as Kepler's laws improves the capability to predict and analyze. As every physics student is taught, Kepler's laws themselves can be derived from Newton's laws of motion and the law of universal gravitation. Newton's laws are considered more fundamental since they help in predicting not only the movement of planets in the orbits but also the motion of cannonballs. Thus one perceives a continuous change from simple extrapolation of data to extremely fundamental laws of nature which are employed in increasing the ability to "know" about any event, in this case the sunrise.

The above examples are from physics. There are many other subjects which will be discussed. The key however is that as one tries

to "know more" one uses more mathematics and logic. The statement of Lord Kelvin "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind" is often challenged by social scientists as an attempt to thrust the model of physics on other areas where it is not useful. We do not directly enter this philosophical argument. One need not go beyond everyday life to notice that using numbers improves our knowledge. "A heap of mangoes" conveys less precise knowledge than "54 mangoes" or "10 kg of mangoes". Given that use of numbers and logic can lead to more and more refined predictions, how does one evaluate if the conclusions are justified. As we shall see it is very possible to overuse mathematics and logic. Feynman had a wonderful word for this, GIGO (Garbage In, Garbage Out). He was referring to highly sophisticated computer models that were being used without proper justification in physics. Since the human brain can also be considered a computer, one can justifiably wonder if human efforts to use mathematics can also be an example of GIGO. It is all very nice to come up with major principles or laws, but how to be reasonably sure that they are right and that they are usable?

Making this argument is the easy part. To "know" the limitations of a given set of observations and the conclusions that can be drawn from them is far from trivial. We consider as a first example the first law of Kepler which states that the path followed by every planet is an ellipse and that the sun is at the focus. An ellipse is not simply a distorted circle. To draw an ellipse a long string passing through a ring is attached to two fixed pins as shown in figure I.1. When a pencil placed in the ring is rotated, keeping the string stretched, one gets the ellipse with the two pins being the foci. Just as all points

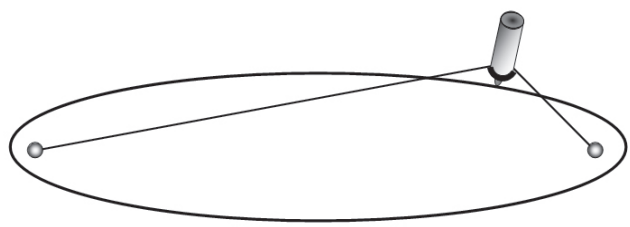

Figure I.1 Drawing an ellipse

on the circle are at equal distance from the center, all points on an ellipse are such that the sum of the distances between the two pins, along the length of the string is constant. The first law, despite the apparent use of descriptive language is a very precise statement about the position of the planet with respect to the sun. Both the sun and the planet actually move in elliptical orbits around the point called the center of mass which is at the focus. In the case of the solar system, the center of mass is very close to the center of the sun itself since the mass of the planet is very small compared to that of sun. So the approximate statement that sun is at the focus is justified. When we consider the movement of two bodies of comparable mass such as a pair of stars, the individual elliptical motions can be detected.

Now consider the results shown in figure I.2. In this picture, actual observations made on the movement of a binary star have been traced removing some details which are not relevant for the present discussion. Clearly the two stars are not moving around a common focus as expected from Kepler's law. A simple minded conclusion

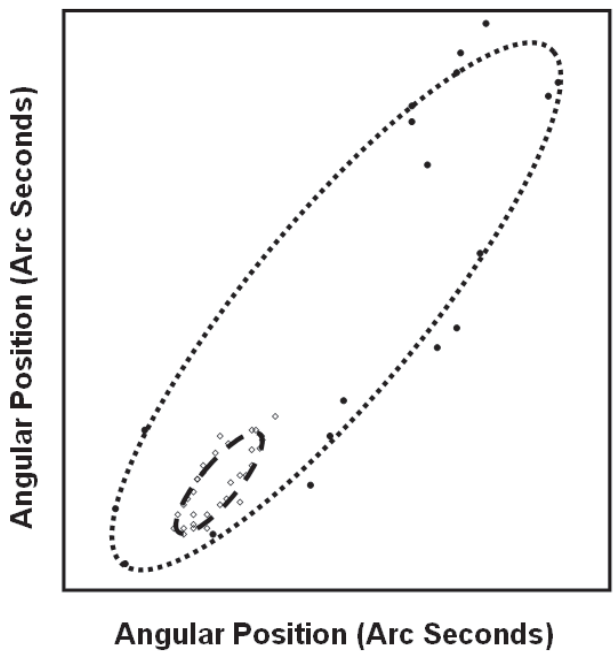

Fig. I.2 Observations show a binary star that does not appear to revolve around the common focus

would be to jump with joy at having found a great astronomical observation that has disproved Kepler's laws and dream of glory. However, the accepted conclusion is that the plane of the ellipse is not perpendicular to the line of observation but is at an angle. Consequently the stars appear not to move around the common focus.

The decision to retain belief in Kepler's law despite such drastic visible evidence is in contrast to the acceptance of Einstein's theory of general relativity. One piece of evidence supporting the theory is that Mercury does not exactly retrace the same path each time it orbits the sun, but swings around over time as shown in figure I.3. The perihelion, the point on its orbit when Mercury is closest to the sun advances with successive revolutions. It is shown greatly exaggerated in the figure. The actual advance is only 43 seconds of arc (less than one thousandth of a right angle) per century. Thus a deviation that is almost impossible to observe is accepted as the basis for a revolutionary revision of the laws of physics. This can be contrasted with the visible disagreement with Kepler's law in the case of the binary star which is simply reasoned away. The hope is that the present discussion enables one to understand the basis for this choice of revising physics (based on the Mercury observations) and retaining Kepler's laws (despite the binary star observations).

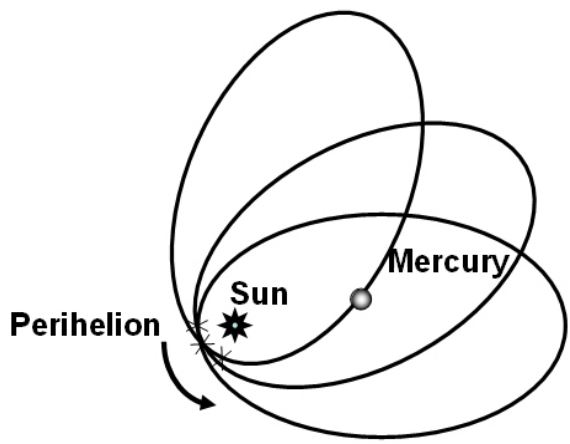

Figure I.3 Exaggerated depiction of the advancement of the perihelion of planet Mercury

Just as disproving Einstein is a major ambition of many "non-physicists" so is the idea of perpetual motion and the resultant availability of "free" energy. We consider another example from Feynman to see how difficult correlating our acceptance of the law of conservation of energy with a specific claim of perpetual motion actually is. Imagine a toothed wheel, with vertical edges in one direction and ramp like edges in the other direction provided with a stopper held down by a spring as shown in figure I.4. The wheel cannot turn in one direction since the vertical edges get blocked by the stopper. In the other direction, the ramp like edge manages to lift the stopper and the wheel can rotate. This detail is shown in the inset. The wheel is connected with four vanes as shown in figure and enclosed in a box of gas. Air consists of molecules which are always in perpetual motion which bombard the vanes and push them. When pushed in one direction the wheel moves but in the opposite direction it does not move. Theoretically one expects perpetual motion in one direction and extraction of energy due to the movement of the central flywheel. Thus one is able to perform work or extract energy from a system without any visible source of energy. As in the earlier example, one can jump with joy at disproving fundamental physics. Most arguments in the free environment of the Internet are similar.

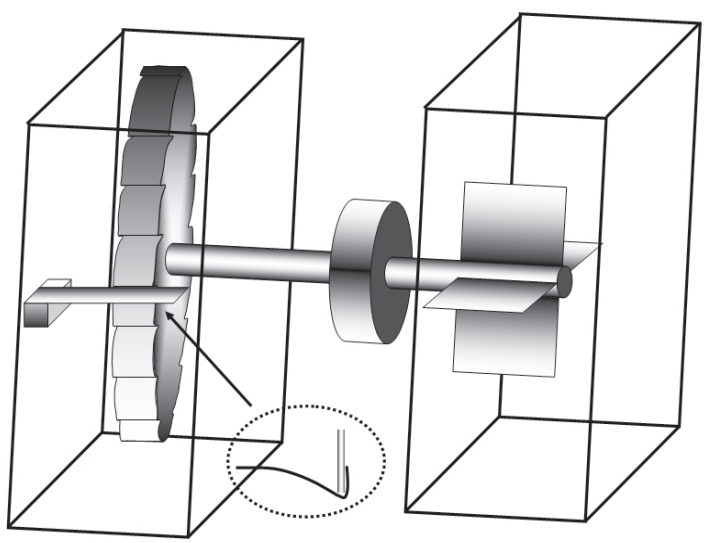

Figure I.4 Explaining the operation of a simple toothed wheel coupled to vanes requires complex physics arguments

Accepting that there is something wrong since this violates the second law of thermodynamics or the law of conservation of energy is very easy. But the idea of science is not to accept the rules from authority. Accepting authority may be good enough if one is invited to join a venture capital fund to invest in a scheme like this. However, accepting by authority is not intellectually satisfying. Identifying the actual mistake in the argument is very difficult. It is however possible to convince oneself that the scheme will not deliver any energy. Consider that there is no friction. When the gas molecules lift the stopper over one tooth, the stopper drops back on to the shaft. Since there is no friction, it will keep bouncing on the shaft. If it is moving up and down continuously the wheel can rotate in the reverse direction also. So for getting one directional motion, friction is needed which heats the mechanism. Thus one realizes that friction and consequent heating is necessary to ensure that the wheel rotates only in one direction. If the spring and stopper are hot enough, they will keep bouncing or vibrating. Once again the wheel can move in both directions. Moreover, if the wheel is hotter than the vanes, the vibration of the spring loaded mechanism pushes the ramp like edge and causes the wheel to moves in the a direction opposite to the one it would move if the vanes transfer momentum to the shaft. The vanes can be then thought of as the damping or source of friction. If there is heat exchange between the vanes and the spring mechanism, which is the case in practice, the wheel will not rotate in either direction on the average. Thus the original idea, that through the thought experiment, one has created a means of extracting unlimited energy is not true.

Examples such as these can require significant time even for a professor of physics to understand and apply the principles of physics correctly. The faculty of Princeton University probably appreciated this issue when Feynman as a graduate student revived the problem of a reverse lawn sprinkler. A lawn sprinkler consists of an S shaped pipe on a pivot. Water comes out at right angles to the axis and makes the S shaped pipe to back away from the outgoing water. Consider a sprinkler submerged in water. Water is sucked into the sprinkler. In other words, the direction of water is reversed. (figure I.5) What is the direction in which the sprinkler would turn? It is logical to argue that due to suction and the law of conservation of momentum, it will turn towards incoming water. It is also possible to

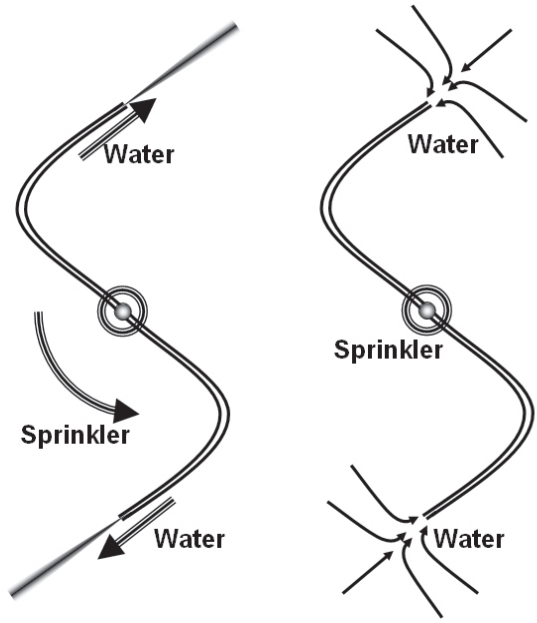

Figure I.5 Another simple apparatus that requires careful analysis, a reverse lawn sprinkler

argue that the centrifugal force of the water is the same whatever the direction of motion of water be and thus the direction motion of the sprinkler should not change.

While Feynman chose not to publicize the correct result even in his memoirs, it is now well accepted that the inverse sprinkler would experience a momentary movement towards the water as the water begins to flow and then stops as a steady flow of water is maintained. While there is a specific direction and hence momentum to the outgoing water which is balanced by the moving sprinkler, there is no such momentum in the reverse sprinkler during steady state. This is indicated by the arrows showing the water direction in the picture. The key point for the present discussion is that even in 1990 the American Journal of Physics (devoted to teaching physics) had received many conflicting analyses of the problem. Apparently "knowing" something very simple even in the hard science of physics is not very easy. The ambitious program of this book is to explore these issues, ranging from justification of induction in the context of fundamental laws to the use and misuse of statistics. One thing is obvious. There can never be a mechanical procedure or a simple com-

puter program to decide how well we know anything. However, it may be possible to learn enough to evaluate "how well do we know it" once "it" has been specified. The answer will not be a binary true or false but one can get some degree of confidence in the conclusions.

1.2 What is not being attempted

It is a huge surprise that improvement in knowledge is possible for human beings. We all seem to be able to understand and use mathematics. Mathematics at even elementary school includes geometry and parallel lines. However, human perception about the mathematical concepts is very unreliable. For example consider the optical illusions in figure I.6 which demonstrate the limitations of visual perception. In the first image, we perceive parallel lines as non-parallel. We can apparently define and understand parallel lines but cannot actually perceive these lines reliably. In the second image {fig.I.6(b)} we perceive a perfect circle as distorted. Using a straight edge or a coin of suitable size we can easily confirm that the perception is wrong but the wrong perception persists. Despite such illusions, we have managed to create elaborate geometry with logical demonstrations. Defining a circle and even determining the ratio of its circumference to diameter to the millionth digit has been achieved. That we have developed geometry is itself a source of wonder.

Almost no one has the capacity to count to ten thousand reliably without making mistakes. This is the typical experience when one handles old fashioned currency notes. Yet we seem to be able to

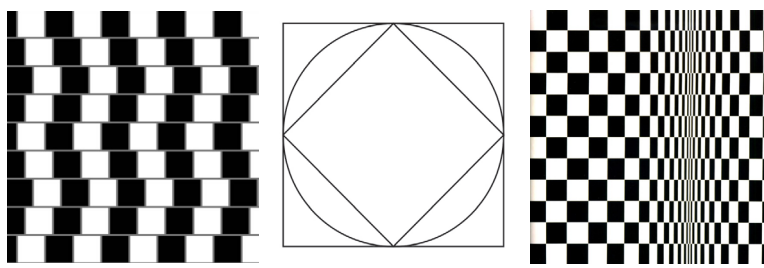

Figure I.6. Examples of optical illusions (a) Parallel lines that appear bent, (b) Circle that appears distorted, (c) Static picture that appears moving

make reliable statements about numbers with an infinite number of digits. It is perhaps important to ask "how well we know" geometry or arithmetic and what we mean by "know" in this context but we shall not discuss this.

Illusions persist beyond the geometric figures. The stationary pattern shown in the figure.I.6(c) creates an illusion of movement. Other examples of our imperfect perceptions are also well known. We perceive the color to be different depending on the background. The art of the magician depends on the ability to create an illusion for the audience. The cinema and the television programs create an illusion based on the limited speed of human eye perception. We perceive as continuous motion, a series of still pictures. Unlike many fishes, a human being cannot directly sense an electric field. Unlike a dog, a human being is neither extremely sensitive to odors nor to ultrasound.

We can be tricked to perceive a change in the color and movement of objects even before the movement has taken place. When two lights side by side are switched on alternately, there is a perception that the light is moving, an illusion that is extensively used in decorating Christmas trees. If two adjacent lamps are of different color, the change in color is perceived midway between the two lamps. The mind has perceived change even before the second lamp has switched on!

When a group of people were asked to count the number times a basketball was passed between players in a video clip, they did not even perceive a person dressed in a black gorilla suit walk onto the court, thumping his chest for about nine seconds and walk off. It is common for an individual immersed in thought to drive on a well known road and be completely unconscious of the traffic through which he has driven.

The feeling of certainty we have when we know something comes from sources beyond our control and knowledge. Certainty is a mental sensation rather than evidence of fact. Because this feeling of knowing seems like confirmation of knowledge, we tend to think of it as a product of reason. But an increasing body of evidence suggests that feelings such as certainty stem from primitive areas of

the brain and are independent of active conscious reflection and reasoning. Not only is the feeling of knowing extremely subtle, rational decision making seems to be very difficult for human minds. There have been extensive investigations of human ability to take rational decisions which show that expectations, emotions, social norms, and other invisible seemingly illogical forces skew our reasoning abilities. More importantly these mistakes in identifying rational solutions appear to be repeatable and can even be predicted. All this makes it extremely difficult to assess how we can know and creates a doubt. "Do we know anything at all?"

Similarly, human use of language itself is very strange. A dictionary defines the meaning of a word in terms of another word which is also in the dictionary. In spite of this circularity, we still communicate and can translate from one language to another. Alternate meanings of the word "know" exist. Statements such as "I know physics", "I know my own mind", "I know something bad (or good) is going to happen" are often made and felt to be meaningful in human conversation. In the present discussion, we ignore the limitations of human perception with respect to both observing the world (the illusions) and the inability to find rational solutions by themselves. We are looking at the simplest methods of knowing that have been with us for a long time. These have been found to be communicable to most humans. We are not concentrating on the beautiful and exciting current research in the fields of cognitive science and evolutionary biology.

1.3 How is it being presented

We begin by discussing how best can we know anything? There is a general feeling that using numbers is a way of qualifying statements and that use of numbers makes it more precise. In the first part of the book we look at certain selected aspects of mathematics. We emphasize that mathematics consists of an intimate mixture of things that are known perfectly and things that are completely unknown. Close association between the known and unknown is observed even in familiar areas of mathematics such as numbers and geometry. This is also true of unfamiliar and complex areas like deterministic chaos. Some of the beauty in this close association acts

as a precursor to the discussion of practical issues later. We also look at the mathematics underlying guessing and some simple problems in probability. Once again there is this association of the perfectly known and perfectly unknown. The toss of a coin for example has only two possible outcomes, a head and a tail. But the outcome of a specific toss is purely random. Statistical description of the random nature of the outcome shows up several unexpected and beautiful results. Even though mathematical quantification carries an image of perfect knowledge, any practical number has an uncertainty associated with it. This uncertainty has some close parallels with the randomness associated with coin tossing. We thus look at the way in which numbers with a degree of uncertainty can be compared and related. This completes the introduction to mathematics and provides an insight into the question how mathematical knowledge itself limits our effort to "know".

The second part discusses the use of the mathematical methods in physics. This choice is deliberate and extremely important. The key issue that is emphasized in this brief summary is how confidence in various aspects of physics varies. Despite the dominance of the basic or fundamental theories in popular imagination, physics knowledge includes empirical functional descriptions and even statistical correlations. This is clearly emphasized by the description of a MOSFET as an example. Thus the fundamental theories or concepts like the conservation of energy in the example of the toothed wheel described above are always true and do constrain the rest of physics and even areas like chemistry, biology and medicine. However, this constraint cannot be taken as a mechanical or automated procedure for evaluating these sciences, nor is it employed as such in advanced scientific research.

This subtle relationship is then employed in the third part to evaluate many other scientific areas ranging from medicine to climate change and economics. The emphasis is to compare the subtle relationship in the so called hard sciences with the rigid and usually unjustified extrapolation of fundamental concepts in other areas of human knowledge. The labeling of these ideas as ideologies rather than scientific theories is then justified and the limits of what science can and cannot deliver are delineated.

While there is a case to be made for knowledge for knowledge's sake, an examination of the question of how well we know anything is ultimately to evaluate the role of human knowledge in guiding human action. The final part takes up this discussion as a conclusion to the earlier description. The conclusions are buttressed by some arm chair philosophical arguments and a reference to a planned companion monograph that will seek to explore areas not covered by this exposition of science, namely religion.

1.4 What is not on offer

The whole point about trying to evaluate how well we know is to avoid taking things on authority. There is no way of decomposing this into a series of simple rules or commandments that can be automatically implemented. Naturally one is not being offered. The book seeks to provide a knowledge base and this cannot be reconciled to light reading for pleasure. Every effort has been made to make the discussion understandable without any prerequisite knowledge but that does not mean effortless understanding. Knowledge is never obtained without effort. The first two parts in the book and in particular the discussions of "chi squared test", "T-TEST" and "unification of theories in physics" could be difficult reading for those with limited previous exposure to mathematics and physics. A clear difference between the first two parts and the rest may also be perceived by the reader. While the discussions in the first two parts are extremely critical for the discussions that follow, the link between them is very subtle. One hopes that in the spirit of the offering, the reader does not leapfrog these sections. However, the book itself is understandable with such leapfrogging if one is not willing to put in the effort to critically examine the basis of some claims in later chapters. Einstein said that "A scientific theory should be as simple as possible, but no simpler". This applies to this modest scientific description.

Part One

Learning From Mathematics

The three R's, taught to every young school student are Reading, wRiting and aRithmatic. If we consider reading and writing as an extension of the natural human capability to speak, mathematics in the form of arithmetic is the first thing that is deliberately taught. Mathematics can be decomposed into a series of extremely small logical steps. Consequently it has been quite easy to make digital computers that perform billions of such mathematical operations in a second. However this very process of decomposing into small automatic steps makes it extremely easy for humans to implement them without any understanding of the limitations of the process. When we explore limitations in applying mathematics as a prelude to understanding how well we know anything, amusing mistakes made by small children (and some adults) spring to mind. It would not be unusual to find a small child make the following addition, 13 + 13 + 13 + 13 + 13 + 13 + 13 = 28. It emerges by simple counting. First add all the 3's to get 21 and then count all the 1's. The mistake of not recognizing the difference between the units and tens is quite easy to make. At a slightly higher level, a mathematics teacher often has to point out to students that while $5\times(2+3) = 5\times2 + 5\times3$, it is silly to perform $5+(2\times3)$ as $5+2 \times 5+3$.

At a more fundamental level, consider the historical development of geometry. Many societies had recognized the truth of many simple geometrical theorems including the Pythagoras theorem. However the

Greeks uniquely converted the study of geometry into a series of logical steps starting with simple axioms; (i) A straight line can be drawn between any two points, (ii) A straight line can be extended indefinitely, (iii) A circle with any radius can be drawn around any point, (iv) All right angles are equal and (v) Given any straight line and a point not on it, there exists one and only one straight line which passes through that point and never intersects the first line, no matter how far they are extended. The last of these, famous as Euclid's fifth had kept great mathematicians busy over more than a millennium trying to derive it from the earlier four. It was only the development of non-Euclidian geometry in the late 19th century that ended this quest.

Obviously, mathematics is a curious subject. It appears to be most logical and few would doubt simple mathematical statements, making them great examples of "knowing". The examples above have been cited to show that both the beginner and the expert can still make mistakes. Despite this drawback, an attempt is made in the following four chapters to understand basic mathematics and its role in quantification of knowledge.

———————

II

HOW BEST CAN WE KNOW ANYTHING?

II.1 What counting implies

Counting is at the heart of mathematics and mathematics at the heart of our discussion on how well we know. Very few however appreciate the complete logic behind the process of counting and all the consequences of counting. The use of numbers is universal. All humans seem to be able to understand the similarity between two oranges and two apples. It is as easy to decide that there is "twoness" common between two oranges and two apples as it is to identify the common "redness" of an apple and of blood. The numeral 2 is thus used as a description or an adjective. Why it is so easy for children to learn this is not at all understood properly. It has been suggested that even small babies possess knowledge of numbers. The experiments performed are quite interesting. Two objects in the field of view of the baby slowly move behind a screen. Then the screen is removed to disclose three, two or one objects. The babies exhibit more interest (stare for more time) when the addition is wrong and either one or three objects are disclosed. The baby seems to be "surprised" by such a result rather than when the screen is removed disclosing only two objects. While the actual meaning of these experiments can be debated, children are perhaps capable of counting intuitively.

Some animals also seem to possess a rudimentary sense of numbers. If a nest contains four eggs, when two are removed the bird generally deserts the nest but not if only one is taken. It is claimed that the bird can distinguish two from three. On the other hand there are reports that languages of some primitive tribes, for example the Bushmen of South Africa, do not have words for numbers other than one two and many. This is rather surprising since they do have names for dozens of animal and plant species. It is important to realize that possessing a number sense does not require the concept of counting. This number sense or the "cardinal" value of a number is practically obtained by matching or tallying every object of one collection with an object of a second collection. This differs from the "ordinal" nature of counting. When counting, we assign to every object in the collection, a term (the numeral) in ordered succession. Thus counting involves the knowledge that the number three follows number two and is followed by number four. This ordered succession permits one to go from one number to its successor.

Counting is the simplest of all achievements. However, it is not merely quantification. Counting involves the process of increment by unity. The apparent simplicity of the counting process and its near universal applicability seems to suggest that counting and simple arithmetic addition represent ideal or perfect knowledge. They do in the sense that except in jokes or poetry no one doubts that one and one make two. What is not so apparent is the reality that the process of counting implicitly assumes the existence of an entity called "infinity", an entity which is a number but one which is its own successor.

That counting implies the existence of infinity can be made clear to even small children. Henry Bethe describes how as a ten year old he was able to grasp the concept thanks to a simple question. "You know there are twice as many numbers as numbers"? After all if we name a number we can also name one twice as large. Thus a number still larger, for example ten times the original can always be named. Thus, there are numbers larger than any given number and this tendency of numbers to grow without limit is the concept of infinity. If the process of adding is limited to some largest number, we immediately see the emergence of large number of questions

without answers. Consider that the limit is 50,000. Then we are able to perform counting and add 25,000 to 25,000 but a large number of additions like 45,000 and 45,000 become undefined though the number 45,000 is itself defined. Rather than create so much confusion, it is simpler to admit that this process can in principle be carried on without limit.

The reason for stressing this is not to discuss the nature of mathematical truths. In the present context we desire to be practical. The discussion highlights the problem with any ideal knowledge. Immediately, things which are quite unknown get revealed. In this example, the perfect knowledge that two and two makes four implies the existence of a not very familiar concept called infinity.

II.2 What simple logic leads to

While counting with numbers we realize that the difference between successive numbers is always unity. Thus the difference between 5 and 6 is the same as that between 56 and 57. At a slightly higher level of mathematics numbers such as 1/2 and 1/4 are introduced. The beauty about these numbers is that we can always identify a new number which is smaller than one and larger than the other. For example 1/3 is smaller than 1/2 but larger than 1/4. Once again 3/10 is smaller than 1/3 but larger than 1/4. There does not seem to be any limit to the process. Given two such "rational numbers" m/n and k/l, we can always find another rational number between the two. In fact we have an infinite number of rational numbers between the two.

Ancient mathematicians in most civilizations were probably aware of this fact though they may not have been able or willing to articulate the infinite. They also were aware of some peculiar numbers which obeyed a startling property. Consider the most common example, (3,4,5), which obeys (3×3) + (4×4) = (5×5). We can also verify that (1/3×1/3) + (1/4×1/4) = (5/12×5/12). In view of these examples, it is a natural question to ask if the sum of the squares of every two rational numbers is also the square of another rational number. For example what are the numbers m and n such that (1×1) + (1×1) = 2 = (m/n×m/n)? We can easily verify that (11/8×11/8) is

slightly less than 2 but (13/9×13/9) is slightly larger than 2. As mentioned above there are an infinite number of rational numbers between 11/8 and 13/9. It is surely reasonable to imagine that one of these, multiplied by itself equals 2 (or 2/1).

It is here that the first idea of a theorem in mathematics emerges. A theorem is a statement which is true whatever be the circumstances. Those with exposure to geometry would immediately have seen the correspondence between the above numbers and the Pythagoras theorem. To know a large number examples of such numbers, called Pythagorean Triplets, is not proving the theorem. It is necessary to provide a series of steps each of which cannot be negated and which confirm that the theorem is true for all the triangles one can visualize. In a similar way it is necessary to logically prove or disprove the above conjecture that some rational number between 11/8 and 13/9 multiplied by itself results in 2 (or 2/1).

Assume that the number we need is m/n. One of the two integers m or n must be odd. If they are both even we can divide both by 2 and get new numbers at least one of which is odd. Now $2 = m/n \times m/n$ or $2 \times n^2 = m^2$. Thus m cannot be the odd number since the square of an odd number cannot be even. So m is even and $m = 2 \times k$ for some k. We now have $2 \times n^2 = m^2 = (2 \times k)^2 = 4\,k^2$ or $n^2 = 2k^2$ or n is also even. Thus we contradict our original assumption that either m or n is odd. This in turn means that none of the infinite rational numbers available would satisfy the requirement that it, (called square root of 2 or $\sqrt{2}$) multiplied by itself results in 2 (or 2/1). Such numbers which cannot be written as a ratio of two numbers are called irrational.

In the modern decimal representation, we all know that a rational number is represented by a decimal which repeats itself. The easiest examples being 1/3 = 0.3333$\underline{3}$ or 1/11= 0.0909$\underline{09}$ or 1/13 = 0.076923076923$\underline{076923}$0. The repeating portion is identified by underlining it. The irrational number $\sqrt{2}$ = 1.41421356237309504880 16887242097… The numbers continue to infinity and there will be no repetition. The use of a few steps of logic has resulted in a number which is known to be unknown! It is known that the sequence of numbers stretching to infinity has no repetitions. This is a bit more complex than knowing that one can count to infinity.

II.3 Knowledge within limits

The relationship between the diameter and circumference of a circle has been investigated in most ancient civilizations that have used a wheel. This ratio called pi(π) has been given approximate values in all cultures. These range from a poetic 3.0 to an amazingly accurate 256/81. As with square root of 2, it is now known that π is also not rational. Any rational number can only be an approximation. Perhaps the ancient approximate formulae such as from the Egyptian Rhind Papyrus (~1650BC), that the area of a circle of 9 units diameter is the same as that of a square of 8 units a side, emerged from experimental measurements. In India, Aryabhata (~500 AD) approximated pi to 3.1416 and may have been one of the first people to recognize that pi is an irrational number. Modern calculations have been performed to more than 40 million digits.

However, a logical method for placing limits on its value has been known since Archimedes of Syracuse (~240BC). He began by

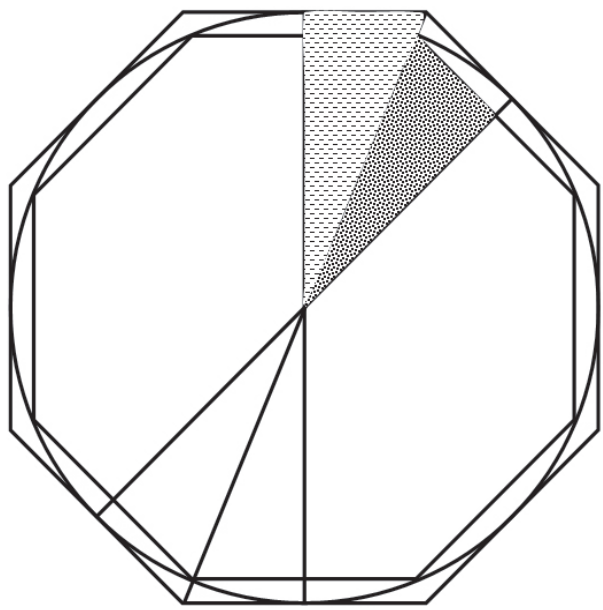

Figure II.1 Archimedes determined the limits on the value of π by drawing regular polygons inside and outside the circle.

stating that the area of a circle must be larger than the area of its inscribed polygon and smaller than that of the circumscribed polygon (see figure II.1 where octagons (eight sides) have been shown). The areas of regular polygons (those with equal sides and equal angles) can be determined easily using geometry. In the right angled triangles that have been shaded in the figure, the angles are known since there are sixteen equal angles in the center. Half the side of the polygon is then related to the radius, (the hypotenuse or long side of the right angled triangle) through simple trigonometry. By using polygons with larger and larger number of sides, we can get upper and lower limits for the value of π. Using polygons with 96 sides, Archimedes gave the limits as $22/7 < \pi < 223/71$. Here is thus another mathematical curiosity. π is a finite number that can only be approximately calculated using the best of mathematical methods.

II.4 Limits to induction

The philosophical argument against induction was briefly alluded to in the introduction with the rhetorical question, "how can the observation of sunrise over the previous million days be of use in knowing whether sun will rise tomorrow?" While evaluating how best we can know anything, it is perhaps relevant to see how much perfect knowledge in the spirit of a theorem, can help in answering any related problem.

Prime numbers are those numbers greater than 1 which are divisible only by 1 and the number itself. If we consider the number 13, the definition requires that when 13 is divided by any of (2,3,4,5,6,7,8,9,10,11,12), the remainder will not be zero. More formally, these are not "factors" of 13. Determining the first few prime numbers is easy. But as the numbers become large, the number of possible factors becomes very large. It becomes more and more likely that one among these is actually a factor and that the number is not a prime number. Is the list of prime numbers finite?

An elegant proof from the Greeks proved that the list cannot be finite. Consider any finite list of prime numbers $p_1, p_2, ..., p_r$. Let $P = 1 + (p_1 \times p_2 \times ... p_r)$. Now P is either a prime number or it is not. If it is a prime number, it is larger than the largest prime number in the list but

it is not in the list showing that the finite list of prime numbers is not complete and correct. If P is not a prime number, then it is divisible by some prime number. Let us say p. If p is in the given list, p would have to divide 1, which is impossible. The number p is thus a new prime number. Either way, the original list was incomplete. Here is one more piece of knowledge regarding the infinite. The number of primes is obviously smaller than number of numbers since the later includes both prime and composite numbers. But we now find that we count numbers to infinity we count primes also to infinity.

If the list of prime numbers is carefully examined, a second list can be made. These are called twin primes since these are two primes separated by 2. For example, (5,7), (17,19) and (881,883) are twin primes. The largest known twin primes as of 2009 have 1,00,355 digits. The real surprise is that the conjecture of there being an infinite number of twin primes has not so far been proved. Apparently the proof that there are an infinite number of prime numbers does not help us to device logic for what appears to be a very closely related problem.

II.5 Surprising results of repetitive mathematics

All numbers are obtained by successive addition of unity. This is the simplest mathematical operation for obtaining a new number from the previous one. When the same mathematical procedure is repeated again and again with the new number it is called an iterative procedure. If slightly more complex mathematical operations rather than adding unity are used, the results are quite surprising. As an example consider the following. Start with an integer N. If N is even, the next number in the sequence is N / 2. If N is odd, the next number in the sequence is $(3 \times N) + 1$. Starting with 12 leads to 12 6 3 10 5 16 8 4 2 1 4 2 1 ... Starting with 909 gives 909 2726 1364 682 341 1024 512 256 128 64 32 16 8 4 2 1 4 2 1…

Both finally enter the endless loop 4, 2, 1. These are called hailstone numbers. Their value resembles the formation of hailstones. As small water droplets move up in the atmosphere with the wind, they grow due to addition of more water and dust. They loose heat to environment since the temperature at higher altitudes is lower. They

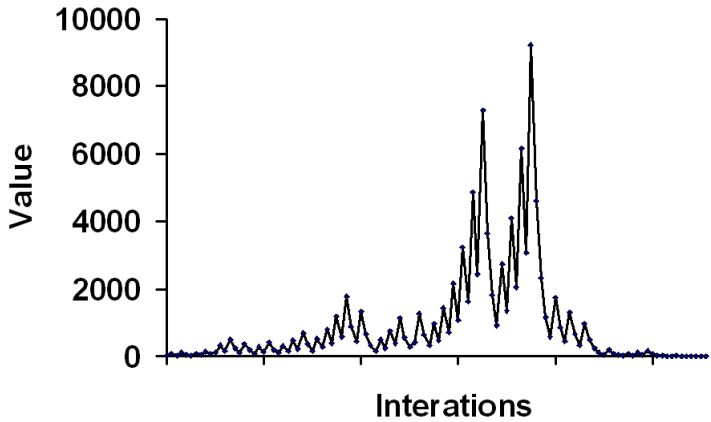

Figure II.2 The hailstone number formed by starting with 27

soon freeze into particles of ice. As the pieces of ice grow, they cannot easily move up with wind and eventually gravity is large enough to bring them to earth as a hailstone. When the calculations mentioned above are performed on different starting numbers, the numbers resemble the hailstones. They grow large before crashing down to the 4 2 1 cycle. A graphical representation of one hailstone number starting with number 27 is shown in figure II.2. It is conjectured that all sequences with any starting number finally fall back to 4,2,1. But this conjecture has not been proved so far. Using high speed computers it has been shown that all starting numbers up to one million do end in the 4 2 1 loop. The beautiful structure of these hailstone number sequences and the eventual crashing into the 4 2 1 cycle emerge from the simple process of multiplication and division.

II.6 Order and chaos from iterative mathematics

Consider another simple iterative procedure performed on a number x chosen to be between 0 and 1. A new value of x is obtained as r×x×(1-x) where r is constant. For example if r = 2.7 and x = 0.02, the new value of x called x(next) = 0.02×0.98×2.7 = 0.05292. If we calculate again with x as 0.05292 we obtain another new value 0.135323. The next several values are 0.315928022, 0.583517268, 0.656167138, 0.609151928, 0.642831813. The values rapidly come close to 0.62963. Surprisingly this limiting value depends only on the value of r that is chosen. It does not depend on the initial value of x chosen. Thus one can start with any value of x in the range of 0 to 1

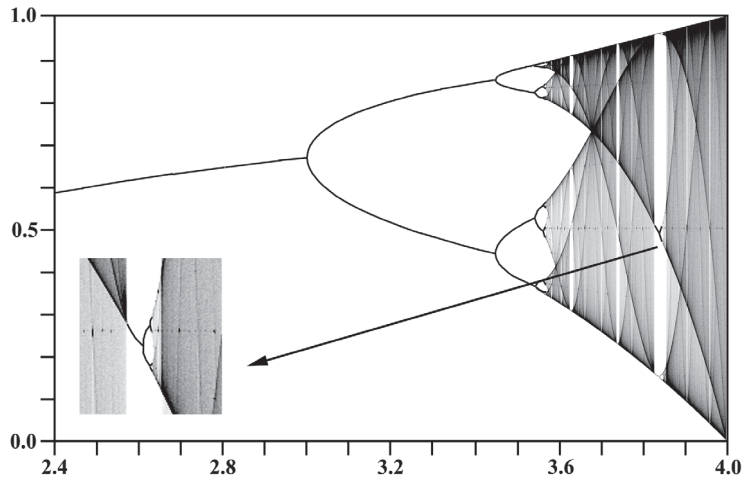

Figure II.3 Figure showing the limiting value of "x" obtained after many iterations for various values of "r". The inset expands a small piece from the main figure to demonstrate self similarity

and obtain the same limiting value after a large number of iterations or repetition for various values of r. Fig II.3 shows a plot of the variation of the limiting value of x with the value of r. Obviously this is mathematics that can only be performed using modern computers. But the process is simple repeated multiplication. The picture shows a very complex behavior. For values of r less than 2.7, a single value of the limiting value is obtained. As r increases further, the limiting value of x oscillates between several values. For example r values between 3 and 3.5 lead to two values of x between which the limiting value oscillates. Actually, whatever be the value of x with which we start, first we approach one of the limiting values. Then each time the process is repeated the value keeps jumping between the two values. At slightly higher values of r the number oscillates between four specific values. Later it is eight values and above 3.7 there is complete randomness and no limiting value. The value will fluctuate randomly over the entire range (0-1) as shown in figure II.3.

This by itself is most unexpected. Repeated multiplications result in either one limiting value, a few values between which the number oscillates or even complete randomness. Which of these three cases is observed depends not on the number with which we started

but on a single parameter r. Systems exhibiting this are called chaotic systems and since the calculation is deterministic (it is simply arithmetic) it is called deterministic chaos.

A more spectacular result is shown in the inset. Here a small region of values of r between 3.8 and 3.9 has been expanded. Surprisingly the whole variation observed initially between r = 0-3.7 is repeated. One sees a single limiting value followed by two and four and so on. This is labeled self similarity. If the mathematics is done accurately enough, this is repeated in each of the small gaps seen in figure II.3 and even in the gaps seen in the inset. Thus there are in principle infinite copies of the initial pattern in the complete diagram. This is called self similarity. There are ways other than doubling of the number of limiting values which lead to chaos in mathematical models. The more relevant issue for the current discussion is the light all this sheds on the question of how best we can know anything. We will have reasons to compare this behavior of simple iterative arithmetic with models employed in many areas of human knowledge.

II.7 The perfectly known and totally unknown

Use of mathematical quantification is at the core of knowing how well we know anything. The usual perception is that mathematical truths are eternal and true beyond human knowledge. The philosophical argument about this is not very germane to the present discussion. The few examples of mathematics discussed here however provide a cautionary warning. The intimate co-existence of the perfectly known and the totally unknown as illustrated by these examples is very striking.

Counting with numbers is a perfect way of knowing. Two and two make four except in humor. On the other hand, the ability to count comes along with the perception of the unknowable infinity. It becomes possible to logically define an infinite number of rational numbers in a finite range and then logically point out that there are numbers which are not rational within the very range. It becomes possible to use numbers which consist of an infinite number of digits where there is no repetitive pattern. It is logically possible to "know"

that there are an infinite number of prime numbers but this knowledge does not help in resolving if there are an infinite number of twin primes. Repetitive use of the basic arithmetic operations of addition, multiplication and division results in beautiful self similar structures that can be appreciated by an artistic mind without understanding the rigor of mathematics. Not merely order and beauty but even chaos can be generated from such iterative procedures.

All this reminds one of Gödel, who proved a logical theorem that there can be theorems in mathematics which cannot be proved to be either true or false. At any instance there are conjectures, statements which are not proved to be wrong by examples. But they have not been proved logically true either. These may be proved in future. Gödel's theorem warns that we may never be able to prove some of them. The limitations of mathematics serve as a caution for deciding practical issues of how well we know anything. Mathematics is an area of knowledge where there is complete intimate coexistence of the perfectly known and the totally unknown. The answer to the question how best we can know anything is to assert that perfect knowledge is possible. No counter example can ever be found for a logical theorem. However, this perfect knowledge is extremely localized. It is like the local regions of well defined behavior inside the sea of chaos observed in figure II.3.

III

HOW TO GUESS WHEN YOU CANNOT KNOW

III.1 The possibility of guessing

In the previous chapter the possibility of representing $\sqrt{2}$ as a rational number was discussed. Before it was logically proved that it was irrational, it appeared reasonable to expect it to be a rational number since an infinite number of rational numbers close to one another were available to represented $\sqrt{2}$ as a ratio of two integers. This was a guess without a logical proof. In the event, the guess turned out to be wrong. Later, the issue of twin prime numbers was also discussed. Currently this is still a conjecture. Based on the evidence that there are infinite number primes and that very large twin primes are known, it may be guessed that there are infinite twin primes but it is not proved so far. In view of Gödel's theorem it may never be. But simpler things than guessing the veracity of mathematical conjectures are relevant to the present discussion.

Consider a toss of coin. This is a very common everyday experience. A coin toss results only in two possibilities, head or tail. Thus before it is tossed, it is certain that either a head or a tail will result but the consequence of a single toss is unpredictable. The

possibilities are certain, in fact absolutely known but the specific knowledge of the consequence of toss is random. Let us consider that the results of a series of four tosses are recorded as H,H,H,H (H=head). Several guesses can be made based on these observations. It is possible that the coin is a special one with head on both sides. It could also have been made in such a way that a toss results in a head more often than a tail though both head and tail are visible on the coin. It could be a fair coin with equal chance for it to land head or tail. If the next two tosses show T,H, making the series (H,H,H,H,T,H) the first of these conjectures is false. To decide between the other two, we need many more trials. If a long sequence such as "H,H,H,H,T,H,T,T,H,T, H,T,H….." is available, it is possible to mathematically confirm if the third conjecture above is true and that the coin is fair. Assuming that this has been done, it is apparent sequences such as HHHH or HTHT present in this series are random, they convey no information. The probabilities of getting a head or tail on the next toss are equal. The HHHH sequence does not increase the probability of a tail on the next toss nor does HTHT of a head on the next toss.

III.2 A surprise regarding ratios and differences

Guessing when the coin is biased is rather easy. If the number of heads and tails are counted for a long series, their ratio will equal 1 if the coin is not biased. If the ratio has a value different from 1, the coin is biased meaning that it favors one outcome. If for example the ratio is observed to be 2, obtaining a head is twice as likely as tail. In practice the ratio of one of the numbers say the number of heads to the total number of tosses is used. This is the probability of obtaining a head. A value of 0.5 means that both the possibilities, heads and tails are equally probable and the coin is labeled as "fair". Other wise the coin is labeled biased.

Sequences, each of 100 trials with coins of different bias are displayed in Table.III.1. If a short sequence such as 10 or 20 trials is selected, there is a high probability that a coin may be considered biased when it is fair or considered fair when it is biased. Short sequences have been highlighted in the first series, that of a fair coin which suggest that the coin is biased. Similarly, in the subsequent series, belonging to biased coins, small sequences that imply a fair

coin have been highlighted. It can be safely concluded that detecting small changes from the ideal ratio of 0.5, to say 0.6 merely by looking at such sequences is almost impossible.

Table III.1

Repeated tosses of fair and biased coins
Head represented by 1, Tail by 0

Number of heads / Total trials	49 / 100	10100011000001111111110 10 01000100101111100111011 10 11001011000000101101001 10001011110101000101000 01
Number of heads / Total trials	28 / 100	10000010010101001010001 00 10000000000110001010000 00 01101101000000001001100 10 00000100000001100001001 01
Number of heads / Total trials	74 / 100	11111111111011111111000 111 00111111111111111001111001 10101011110101110011110 11 01111011111111101010101 00

This is the mathematical way of determining if a coin is fair, by determining the ratio of the number of heads and tails for a large number of trials. Assuming a fair coin, the ratio will approach closer to the value 0.5 as the number is increased. For a typical example, if we consider tossing a coin 10 times, the number of heads will usually vary between 2 and 8. On the other hand if we try the procedure with a 100 tosses, the number of heads will vary over a smaller range. The variation would usually lie between 40 to 60 heads out of 100 trials. Trials with even larger number say 500 will reduce the range even further. A trial of 10 tosses can result in obtaining a ratio of 0.2 but a trial with 100 trials will not result in such large deviations from the true value.

The most important point to note is that the difference between number of heads and number of tails does not decrease even though the ratio will get closer to 0.5. This at first sight appears counter intuitive. In the earlier example, initially the ratio (0.2) was much smaller than the ideal value (0.5). This means that the number of heads was much smaller. When the number of tosses is increased, the

ratio becomes closer to the ideal value which means that the number of heads in the newer part of the series was larger. This does not however guarantee that the number of head and tails are equal or even that the difference between the number of heads and tails is smaller than before.

An easy way of understanding this is to imagine that after 1000 trials there are 510 heads. The ratio is 0.51 and there are 20 heads more than tails. If 100,000 trials are made, the ratio has to be closer to 0.5. It can be 0.5004 which is nearer to 0.5 than 0.51. There are however 800 more heads than tails in the 100,000 trials. Thus, though the number of trials has increased and the ratio is closer to the ideal value, the difference between number of heads and tails has not decreased. This interesting result is the basis for many fallacies that will be discussed when we evaluate how well things are known.

III.3 Creating randomness

It is not very easy to deliberately create a series indistinguishable from the one obtained from a real tossing of coin. Consider two scenarios. One in which a series, HTTHHHHT… is produced by tossing a real coin. The second involves mentally guessing and writing down such a series without actually tossing a coin but imagining that a coin has been tossed. Is it possible to distinguish the two? The answer surprisingly is that they can be distinguished quite efficiently. This exercise is sometimes undertaken in elementary statistics classes where randomly some students are asked to guess such a series and some asked to actually toss the coin. Long stretches such as HHHHHH or TTTTTTT appear in a genuinely random sequence 100 items long. When we deliberately try to create a random series, the maximum length of the regularities is smaller than expected. As a consequence the number of times there is a switch from T to H or H to T is much larger. These key signatures help a trained statistician to distinguish between the real random sequence and one created deliberately. The lesson that emerges from this exercise is that "it is very hard to appear to be random on purpose" and identifying randomness through visual observation is very difficult. A simple sequence HTHTHTHT will immediately reflect the order to anyone. A real sequence does not have that much order but there are several short ordered segments.

III.4 How ordered is a random sequence?

How long must a series of coin tosses be to obtain a series of "n" heads? Obviously there is no fixed number for this. After all we know that HHT, THH, TTTTHH etc. are all possible. In the first example, two consecutive heads were obtained after only two trials, in the third example after six trials. The minimum is obvious. Only two trials are required for getting two heads in a row. But just as it is possible to define the ratio of the number of heads and tails to determine the bias of a coin it is possible to obtain an average number of trials required to obtain a series of "n" heads or tails. It can be expected that this average number should depend on the bias of the coin. A series of four heads for example should be more easily obtained when the coin is biased to give more heads than tails.

Table III.2 shows the average number of trials required to obtain a series of "n" heads for n ranging from 2 to 7 for various values of the ratio of heads to total number of trials, or the bias of the coin. For a fair coin the ratio is 0.5 and the number of trials required to obtain a string of six consecutive heads is little more than 100. The surprising number to be noted is the number of trials required to get a series of four heads when the coin is in fact biased to provide two tails for each head. The bias is 0.33 and once again a bit more than 100 trials are required to obtain this. Obviously it is simply impossible to determine the bias of a coin by simply looking at the series. This is even more serious if the series is rather short. As shall be discussed later this is extremely relevant for the discussions attempted in this book. Real life "knowledge" consists of a small number of repetitive

Table III.2

Average number of tosses required to get 2-7 consecutive heads for fair and biased coins

Bias	Number of Heads					
	2	3	4	5	6	7
0.33	12	40	124	380	1154	3500
0.4	9	24	63	161	405	1015
0.5	6	14	30	62	126	254
0.6	4	9	17	30	51	87
0.67	4	7	12	19	30	47

observations (positive or negative) from which conclusions are drawn about how well we know. The "traditional knowledge" or "cumulative experience" is formed without formal statistical analysis and often based on short sequences of observations. Possibility of such sequences due to random causes must be considered but as shall be discussed in future chapters this is rarely done.

III.5 Identifying randomness

It appears that long strings of heads or tails are possible even with the coins biased against their occurrence. Mere observation of the sequence cannot confirm that the sequence is random. So the veracity of a single number, the bias of the coin as determined by the ratio of heads to tails for a large number of trials can be questioned. For example a repetitive sequence HHTHHT… would also provide the ratio value of 0.67 but the tosses are obviously not random. However it is possible to obtain more information from any given sequence.

To examine an arbitrary sequence for the randomness, several series of a fair coin, each of 100,000 trials are generated. This becomes easy to perform using a computer and the results are plotted in figure III.1, after performing some further mathematical operations. To begin with, the number of heads in a short series of 4 values is counted. The

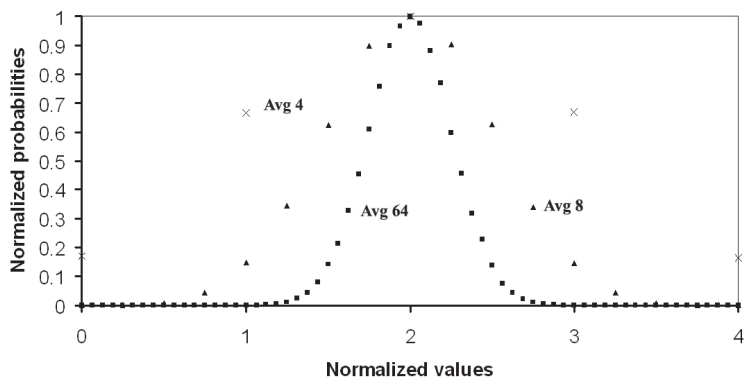

Figure III.1 The Gaussian probability distribution produced by averaging the tosses of a fair coin (head=1, tail=0) over 4, 16 and 64 trials

number can range between 0 if 4 tails are obtained and 4 if 4 heads have been observed. Many such sets of 4 are obtained from the trial data. The frequency of occurrence, the number of times any specific number of heads (0,1,2,3, or 4) has been obtained is determined. The individual frequencies are divided by the maximum value so that they vary between 0 and 1. As mentioned before the probability approaches the ideal value of 0.5 as number of trials is increased. Thus one expects that the curve should reflect a peak at 2 heads and one is observed.

The same procedure is performed for sets of 16 and 64 each on 100,000 independent trials. The values obtained when the number of heads is counted in sets of 16 or 64 have been divided by 4 and 16 respectively so that the values are plotted within the same range of 0-4. This has been done to plot all data on the same graph. The data shows a peak which becomes sharper as the number over which the results are summed is increased. Thus the curve corresponding to summation of 64 trials is a very sharp peak at the center. The peak for summations of 16 trials is not as sharp but still sharper than when averaging is performed over a set of 4.

The shape of this graph, called Gaussian or Normal distribution in scientific circles and a bell curve while describing it to the general public is a source of immense political controversy some of which will be discussed later. The width of the curve reduces as the square root of the number of trials being summed. Thus we obtain an indirect way of knowing if the sequence is random. By summing suitable sized trial sequences we can confirm the expected reduction in width. This observed reduction in width, called the reduction of standard deviation is another very relevant feature of practical knowledge and will be discussed later.

III.6 Comparing randomness

Consider several independent sources for sequences apparently obtained by tossing a coin. To have confidence in the value of the bias of the coin it is necessary to compare the randomness. A simpler method than the one outlined earlier is needed since in practice, the number of trials cannot be increased indefinitely. This situation is exactly similar to many research investigations. In that case the bias

will correspond to a successful or positive research conclusion. It is necessary to confirm that the positive research conclusion is a not consequence of randomness.

This is often accomplished by using the "chi square test". As before, the number of heads obtained in a small group selected from the total trial is considered. If the coin being tossed is fair, and we consider a set of 10 members, the ideal expected value for the number of heads is 5. As the bell curve shows, while 5 is the most probable value, there is reasonably large probability of obtaining a different value. If the first set had a value 2, the difference is 3. The normalized squared difference is calculated as the square of the deviation from the ideal value divided by the idealized expected value, 3×3/5=1.8 in this case. Since the total trial is divided into "n" small sets of 10 each, we have "n" values of this normalized deviation. These are all positive quantities. Squaring ensures that a positive deviation does not cancel a negative deviation. Thus 2 and 8 have deviation -3 and +3 but the same normalized deviation. Squaring also ensures that a large deviation makes a larger contribution to the chi squared than smaller deviations. A deviation of 1 contributes 0.2 but the deviation of 3 contributes much more than three times this value. Normalization, (dividing by the expected value) ensures that the value does not depend on the value of the deviation but only on the deviation as a fraction of the expected value. A large deviation is a signal of an error. The sum of all the normalized deviations for the "n" sets is called the chi squared value.

Now consider the data in Table III.3 generated by tossing a ten sided dice. The probability of getting each of the digits 0-9 must be equal for fair dice. The number of times each digit has been observed in 1000 trials is shown below for two different dice.

Table III.3

Results of 1000 trials of rolling two different simulated 10 sided dice

Digit	0	1	2	3	4	5	6	7	8	9	Total
Number	112	90	87	123	99	86	89	102	98	104	1000
Number	92	95	104	107	89	103	94	105	97	110	1000

The chi squared value for the first series is 11.88 while for the second it is 4.54. This number is converted into a probability using standard theoretical plots showing the probability of occurrence of a given chi squared value. The plots are well known and shown in figure III.2 for different degrees of freedom a term which will be defined presently. Using them, from the value of the chi squared, the probability that the given sequence of values has been generated by a truly random process can be determined. For example, the shaded area for the curve with ten degrees of freedom corresponds to a limit of 0.05. That is, the probability that an event is caused by pure chance is less than 5% for the values of chi squared in the shaded area. If the chi squared value is larger than this, the probability that the data represents pure random chance is higher. This limit is a popular choice for a variety of research problems as will be discussed later. Calculating the chi squared enables comparison of two or more sequences. In the figure the curves are labeled with 1, 6 and 10 which are the degrees of freedom. In the present example there are ten sets and the probabilities are taken from a curve (not actually shown) with 9 degrees of freedom. From such a plot, the probability of the first series in the table being obtained by random process is 0.22 but it is 0.87 for the second. The second is superior. There is a higher probability that the second set of data have been obtained from a "fair" die.

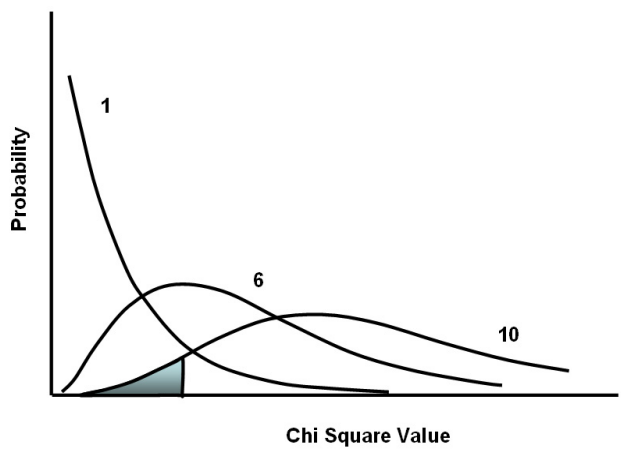

Figure III.2 Probability of random chance being the cause for a given value of chi squared and for different degrees of freedom. Shaded area corresponds to 0.05

The degree of freedom is an interesting concept and denotes the number of independent observations. In the example, the total number of trials 1000 is fixed and known. Once the number of times digits 0-8 are observed is known, the number for digit 9 is already determined. It is obtained by summing the numbers for digits 0-8 and subtracting it from 1000. Thus even though there are 10 sets of data one of it is determined by the others and hence 9 sets can have independent values. Thus the degree of freedom is 9.

By calculating the chi squared value, it is possible to compare two random processes and identify the superior one. More importantly for the problems of interest in the present analysis, the chi squared value gives the probability that the observations are caused by pure randomness. Smaller this probability, more is the confidence that the observed effect or phenomenon is real and not the result of pure randomness. Obviously to have confidence in the observed results, we would like to ensure a very low probability of it being caused by randomness or pure chance. In the example above however, two random processes were being compared and so the one with a higher probability of being random was termed superior.

III.7 Knowing the bias

Just as one would like to compare between two sequences which are claimed to be random, one would also like to know the accuracy with which the bias of a coin can be known. The bias of the coin is ideally known when the ratio of the number of heads and tails has been determined for infinite number of trials. In practice, the bias is determined from a finite sequence of heads and tails but the error can be estimated. If the bias is determined to be p and a sequence of n tosses has been used to determine this bias, a detailed mathematical analysis confirms that the error in the value of p is $\pm 1/\sqrt{n}$. This is true only if the sequence is random. Thus, this estimation of the error can be used only if it has already been confirmed that the sequence is random or in view of the discussion above reasonably random. In view of this, if the sequence consists of 100 tosses of a fair coin, the error is 10%, and it is 1% if the sequence is of 10000 tosses. If the coin is biased, the error is smaller. Assume that the bias of a coin is 0.9 in favor of heads. As was discussed earlier a series of 9 heads out

of 10 is more likely than if the coin were fair. Consequently one can expect that the confidence in the estimate of the bias would to some extent depend on the bias itself. Thus when only a short sequence of trials is available there is a question of how representative this is of the reality. If several sets of "n" trials each are investigated, the randomness alone would ensure that the value of bias calculated from each of these will not be identical.

In practice when these statistical procedures are applied to real life situations, for example while conducting an opinion survey, not only is a small sample selected, the population itself is finite unlike the number of coin tosses which can be extended indefinitely at least in theory. The percentage quoted by an opinion survey is obviously equivalent to the bias of the coin. So determining the error in this is slightly more complicated. However, modified formulae exist which enable a final statement to be made that "p" is the percentage (in the present description the bias) which is known within an error "E" with a given degree of confidence usually 95%. Thus when a value such as 46% is mentioned with a 3% error, the statement means that repeating the survey in the given population will result in a value in the 42-46% range at least 95% of the time.

III.8 Uncertain identification and the base line dependence

While discussing the randomness in the sequence of trials and the limit on the value of the bias, it is assumed that there is no error in the recording of the trials themselves. In other words, if an ideal coin on flipping has turned up "head" it has been identified as "head" and so recorded in the sequence. However, in any practical situation this is not always true. A "head" might have been recorded as a "tail". In the case of a real coin, this for example could be the result of poor illumination. When this happens, the two numbers used to determine the bias of a coin namely the number of heads and number of tails would have errors. A rate of error for wrong identification can be assumed or determined practically but it is not automatically the error in the bias. The same error in wrong identification could create different consequences depending on the bias of the coin. This has

important consequences in practice as many examples which will be discussed later demonstrate.

For the present consider that x% of the heads and tails have been wrongly identified. Thus the actual number of "heads" is obtained by subtracting the number of "tails" wrongly identified as "heads" and adding the number of "heads" wrongly identified as tails. If the error rate for both wrong identifications is the same the result for a fair coin will not change. Since the coin is fair the number of heads wrongly identified as tails and vice versa will be equal. Thus the coin will still be recognized as fair. The error in the estimate of the bias is equal to the error in identification of heads and tails namely x% as assumed above.

On the other hand consider a coin which is severely biased with a bias value of 0.1 for example. This will mean that there are nine times as many "heads" as "tails". Correspondingly the number of heads wrongly identified as tails will be very large compared to the number tails wrongly identified as heads. This is not because the error rate is different but because the number of tails is very small. As a quantitative example, consider that the error in measurement is 5%. In other words, 5% of heads and 5% of tails have been wrongly identified. In the present example with a coin of bias 0.1, there are nine times as many wrong "heads" as wrong "tails". More wrong "heads" get added to the small number of correct "tails" while only a few wrong "tails" get added to the correct heads. The same 5% error in identification would result in 40% error in the measured value of bias. If the error in identification was 10%, the error in determining the bias would have been 80%. Simply accepting the error in the identification of heads and tails with the error in the value of bias has very important consequences for knowing how well we know. The error is called base line fallacy.

While practical examples were not cited in these chapters on mathematics, base line fallacy is so very important that we emphasize this in the language of real life examples that mirror the case of the tossing of coins. Base line fallacy is very relevant when a witness identifies correctly the color of a car in a court of law or when a sophisticated test confirms that the given sample of blood is HIV +.

By performing a series of tests in the laboratory, the error rate in either witness identification or the HIV test can be determined. This corresponds to the error in counting heads and tails or error in identification. However, the confidence in the test, or the error in the actual situation, is equivalent to the error in the bias. As described above this depends on the value of the bias itself.

Consider that in laboratory a test is 99% accurate in detecting earthquakes. In other words whenever there is an earthquake in the next 24 hours, the test gives a positive result 99% of the times and a negative result when there is no earthquake once again 99% of the times. What is the confidence a government should give to the announcement by the scientist that a positive signal has been observed? Does it mean that the chance of an earthquake in the next 24 hours is 99% and immediate action is necessary?

It is in such situations that the base line fallacy becomes most important. Unfortunately for the scientist, the chance of an earthquake is quite low. A typical large earthquake occurs even in areas prone to earthquakes such as Japan or California only once in a decade or so. If the test is repeated a million times in the period, there will be one true positive when there is an earthquake and 10,000 wrong positives due to the 1% error in the test when there is no earthquake making the prediction useless.

Consider a case when the witness in the legal case is tested with a mixture of 50% white and 50% black cars and the person makes a mistake once every 10 identifications. If the same is now duplicated with a mixture consisting of 10 % white and 90% black cars, his ability to identify a white car is much lower. With his 10% chance of error, he would make far more wrong identifications of black cars as white. Thus out of a total of 100 identifications he would identify 9 white cars (correctly) and 9 cars as white when they are black (10% of 90) thus his accuracy of being able to identify a white car is nearly 50% and not 10%. Such dependence on the base value (whether the mixture has 50% or 90% white cars) is extremely important in many different areas of human endeavor as will be shown later. While in recent years there have been concerted efforts to learn and apply these ideas of statistics in various areas where they are

relevant, many still fall into base line fallacy and assume that since the test is accurate to x%, a positive result implies the disease with the same accuracy.

The most important information to be obtained by the present discussion is the emergence of qualified answers for simple questions. Is a given coin or dice "fair"? While a fair coin or dice can be defined, they cannot be proved to be truly fair in practice. The best we can do is to compare two such claims or to give a probability that the given experimental data (the result of tossing a coin for example) is actually "random". Similarly, the bias of the coin can only be known with a little bit of uncertainty which depends on the length of the sequence and the bias of the coin. Finally when determining the "actual bias" of biased coins, the configuration of the test sample itself impacts the result due to base line dependence.

The discussions in the last chapter showed that even with simplest mathematics, perfect mathematical truths coexist with the completely unknown. As was mentioned earlier, the toss of a coin has the perfect certainty that the result would be a head or a tail but without a prior knowledge of which. Detailed analysis of the data generated by such tossing provides several important conclusions. Firstly, guessing if the coin is fair purely by looking at a few results is impossible. Secondly, there is no reason to expect that the number of heads and tails will be equal in due course. Finally the randomness that one expects can only be qualitatively evaluated. These features will be used in determining how well we know anything in real life.

IV
HOW TO COMPARE UNCERTAIN NUMBERS

IV.1 The necessity of uncertain numbers

The previous two chapters introduced uncertainty as an inseparable part of mathematical and logical knowledge. While the sequence of heads and tails is random in principle, it can be quantified to the extent of comparing the randomness of two sequences. While the sequence of digits constituting the decimal representation of $\sqrt{2}$ is random the magnitude of the number itself is not uncertain. It is possible to clearly and logically answer a question if another number is larger or smaller than $\sqrt{2}$. Thus we know that 1.4, 1.41, 1.414, 1.4142 are all smaller than $\sqrt{2}$ and 1.5, 1.42, 1.415, 1.4143 are all larger. So far we have not considered numbers whose magnitude itself is uncertain. This uncertainty is similar to the question addressed earlier regarding the randomness of the sequence of heads and tails. How can it be practically confirmed that a given sequence is random. The same issue of practical realization emerges when a number is considered not abstractly as in mathematical theory but in practice. It has been mentioned in the second chapter that a number can have both a cardinal and an ordinal value. In the first we tally a number against another like a child counting his toy bricks or a man counting

currency notes in a bundle. This process is practical but inherently error prone. By defining an ordinal value of numbers, this error has been ignored in mathematics. Since the book discusses practical matters, we return to cardinal number sense and the errors associated with it. Thus the uncertain numbers used in practice have the features discussed in the previous two chapters namely ordinal value and uncertainty.

Errors in counting are more common than expected. In a game of cricket the umpire uses pebbles to help in counting the six balls that constitute an over. Personal doubts while counting currency notes are a common experience. Once the numbers increase, the chance that a mistake is made will be rather high. Consider that a heap of pebbles are being counted and different attempts at counting by the same or different individuals results is a series of numbers which may or may not be equal. One number in the series may be the hypothetical true number. But there is no mechanism of identifying it correctly.

IV.2 Using the sequence to guess a correct value

One of the simplest approaches to identify the correct value from the sequence would be to claim that the correct number would be the most common. Thus if the series consists of (99,100,97,100,102, 100, 98) there is one reason to choose 100 as the correct number. This has appeared most often. Statistically this is called the "mode". This selection however means that all other trials of counting are rejected. Consider a different series (99,101,97,100,102,101, 98, 99,100,97,99,100,100,98,97). 100 is still the mode since it is been observed four times but it begins to look a bad choice when only 4 out of 15 trials have reported that value.

There are other more troubling possibilities. If one member in the above series is changed we get (99,101,97,100,102,101,97, 99,100,97,99,100,100,98,97). We can see that both 97 and 100 are observed four times each and there is simply no easy way to decide what the correct value has to be. Thus, while the most common observation is a natural choice for the correct answer, it has very little justification as a general choice. The major problem with the choice

of a "mode" is ignoring uncertainty. By selecting mode, the other trials are all ignored and artificially an absolute ordinal value is introduced. The mode does not carry any information about the uncertainty in the counting process.

A second approach is to arrange the numbers in an increasing series. Thus a series (97,97,97,97,98,98,98,99,100,100,100, 100,101,101,102) can be considered. Here there are seven results larger than 99 and seven smaller than 99 resulting in a unique status for this value. A value which has 50% of the trials reporting a smaller value and an equal number reporting a higher value is called the "median". The advantage of the "median" over "mode" is the extra information it provides regarding the uncertainty. However the extra information is "of a somewhat unsatisfactory kind" as in the quotation attributed to Lord Kelvin. The uncertainty cannot be quantified. Examples can be cited where the median is obviously a wrong choice. For example, the above series can be modified as (92,92,94, 95,97,97,98,99,100,101,103,103,104,105,108). This has so much more variability but the median would still be the same.

Since antiquity, three methods of averaging numbers have been used. These provide a "mean" value that is a combination of all the trials. The most familiar is the arithmetic mean where the values are added and divided by the number of trials. Thus the mean value of the trials (98,97,102,103) is 100 not withstanding the fact that 100 has not actually been the observed value in any trial. The other two means, are not as famous. One is called the geometric mean obtained by multiplying the "n" trial values and taking the n^{th} root. So the geometric mean of the above set would be obtained by taking the fourth root of the product of {98,97,102,103}. The geometric mean is 99.97. ($99.97\times99.97\times99.97\times99.97 = 98\times97\times102\times103$). A third mean called the harmonic mean is obtained by calculating the arithmetic mean of (1/98,1/97,1/102,1/104) and taking the reciprocal. In this case we obtain 99.93. The arithmetic mean gives importance to trials which give a very large result while the harmonic mean gives more weight to the smallest values.

In the above example, the arithmetic mean is larger than the geometric mean which is in turn larger than the harmonic mean. This

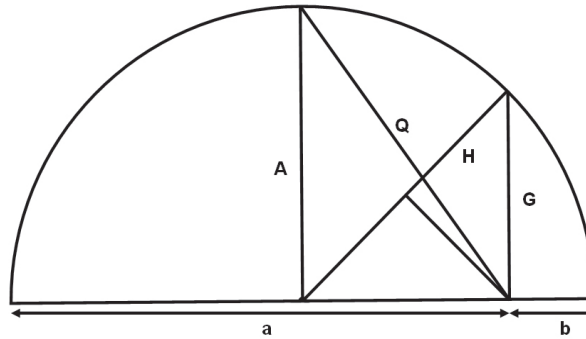

Figure IV.1 Geometric construction of arithmetic, geometric and harmonic means

inequality is always true. The geometric construction developed by the Greeks to define these means is shown in fig. IV.1. The two values a and b are shown as line segments and the semicircle is constructed with a+b as diameter. The radius is thus the arithmetic mean. Q is the hypotenuse of a right angled triangle with sides (a+b)/2 and (a-b)/2. Q is thus $\sqrt{(a^2 + b^2)/2}$ (by Pythagoras theorem) or the root mean square value much used in physics. G, the geometric mean is one side of another right angled triangle. One side of the right angled triangle is (a-b)/2 while the hypotenuse is the radius or (a+b)/2. The radius is the Arithmetic mean A. Thus G the geometric mean is the $\sqrt{((a+b)^2/2 - (a-b)^2/2)}$ or of $\sqrt{ab}$. Obviously it is smaller than the hypotenuse or the arithmetic mean. The harmonic mean H is one side of the right angled triangle with the geometric mean as the hypotenuse and is hence smaller than it. The self explanatory nature of the proof of the relative magnitudes of the various means is quite impressive. It is also apparent that while the arithmetic and the harmonic means will be rational numbers if the trial values are rational numbers, the geometric mean can be an irrational number.

The use of a mathematical process immediately raises one simple question that is quite relevant for the present discussion. Consider the arithmetic mean of three trials (2,3,2), 2.333333. If one were counting pebbles or other objects the meaning of a fractional pebble will be puzzling. Even if the process is extended to other measurements that result in real values rather than integers as with counting, how many decimal digits have to be used? The answer to the question takes us far into practical issues that depend on the specific

details of the measurement. Since no general answer suitable for the present discussion is available, this issue is ignored.

The more important limitation of the mean is that there is still no information on the uncertainty reflected in the sequence of trials. Thus sequence of trials (98,97,102,103) and (76,92,101,131) both give the same value of the arithmetic mean, 100 but the sequences are obviously different. The limitations of the mean are only marginally less than that for the median and mode.

IV.3 Extracting information from the sequence of trials

A single number, the mean, the median or the mode normally cannot provide any information regarding the uncertainty in the sequence of trials. Providing the entire sequence does not help in guessing the correct value. To make practical use of the sequence one has to analyze the reason for the results of different trials being different. We consider the case where the differences represent random errors. Averaging of random numbers described in the earlier chapter suggests that random errors cancel out. There is a probability for the error to be positive or negative and when added these two errors cancel out. This was the reason why the summation of random numbers led to a bell shape curve. A measurement with a random error represents

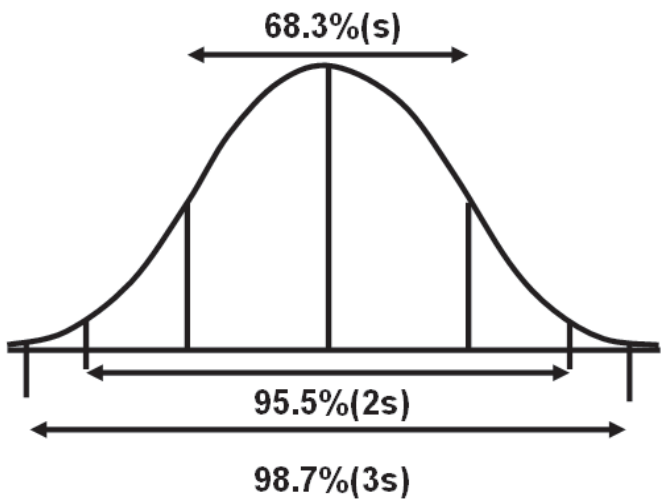

Figure IV.2 The Gaussian distribution and the confidence levels

a random addition to the true value. Consequently, an uncertain number can also be expected to be shaped as a bell curve as shown in figure IV.2. In case the errors are not random, the errors represent wrong experimental procedures and have to be examined on an individual basis just like the number of digits to be used, mentioned above.

When errors are random, the median, the mode and the mean are all equal and this is also the most probable value. The curve shows the probability of obtaining any particular value. According to the plot, the probability of a trial returning a value far from the mean is much smaller than that of obtaining a value nearer to the mean. The standard deviation "s" is calculated for these probabilities. The process is very similar to calculating the chi squared described earlier. The square of deviations from the arithmetic mean are determined. The arithmetic mean of these squared deviations is the variance and its square root is the standard deviation.

When the trial is represented as a Gaussian, the probability for getting a result mean ± s is 68.3%, mean ± 2s is 95.5% and mean ± 3s is 99.7%. This is true only if the errors in trials are truly random. When the errors are truly random, the sequence provides a confidence limit. The standard deviation provides a quantitative confidence level for the results. For example, when the mean is 100 and the standard deviation is 4, the confidence level says that 68 out of 100 trials would result in a value between 96 and 104 (100±4). If the length of the sequence is increased so that there are four times as many trials, the standard deviation will reduce to half the original value. Then 68 out of 100 trials would result in a value between 98 and 102 (100±2).

IV.4 Reducing the random errors

The reduction of random errors due to averaging does not depend on the relative magnitudes of the error and the true value. As the standard deviation increases to a large value, the bell shape curve closely resembles a horizontal straight line. This is exactly similar to the situation in the case of the random numbers. With averaging the true value will emerge out of the random errors even if the magnitude

of the error is a thousand times larger than the mean value. However this benefit will only be realized when the errors are truly random. If the errors are not random, they cannot be expected to cancel out each other in the long run. The consequence would be that there is no true value over which the randomness is superimposed. There is no true value to be identified by this mathematical procedure.

Table IV.1

Numeric data demonstrating the recovery of the signal shown in (a) by averaging 32 copies of (b) resulting in (c)

(a)									
0	0	0	0	**1**	0	0	0	0	0
0	0	0	0	**1**	0	0	0	0	0
0	0	0	0	**1**	0	0	0	0	0
0	0	0	0	**1**	0	0	0	0	0
1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
0	0	0	0	**1**	0	0	0	0	0
0	0	0	0	**1**	0	0	0	0	0
0	0	0	0	**1**	0	0	0	0	0
0	0	0	0	**1**	0	0	0	0	0
0	0	0	0	**1**	0	0	0	0	0

(b)									
1	1	1	0	1	2	-1	2	2	1
-1	0	2	2	1	2	2	-1	2	0
1	2	-1	1	1	0	1	1	0	0
2	2	1	2	3	-1	1	2	-1	1
0	3	2	0	2	0	0	3	1	0
0	0	0	2	3	0	0	-1	-1	1
0	0	1	0	2	-1	-1	1	-1	0
2	0	2	0	1	2	1	1	0	-1
1	1	2	2	0	1	2	0	1	1
1	0	0	0	1	-1	-1	2	-1	0

(c)									
0	0	0	0	**1**	0	0	0	1	0
0	0	0	1	**1**	0	0	0	0	-1
-1	0	0	0	**1**	0	0	0	0	0
0	1	0	0	**1**	0	0	0	-1	0
1	**2**	**1**	**1**	**3**	**1**	**1**	**1**	**1**	**1**
0	1	0	0	**1**	0	0	0	-1	0
0	0	0	-1	**1**	0	0	0	0	0
0	0	0	0	**2**	0	1	0	0	1
0	0	0	0	**1**	0	0	0	0	0
0	0	1	0	**1**	0	0	0	0	0

The reduction in random errors can be visualized using the array of numbers shown in Table IV.1. A pattern is immediately visible to the eye in the set (a). The values forming a cross are all "1" while all the other entries are zero. Now to each number in the array, a random number is added. The random number ranges from -2 to 2. A sample is shown in set (b). The pattern visible in (a) is not visible in (b). The addition of a random number is simply equivalent to adding noise. The "pattern" or the signal in which one is interested is lost due to the presence of noise. Now 32 such sets are generated and average values are calculated for each position, using the corresponding 32 values. The result shown in set (c). Once again the pattern becomes visible though it is not as good as the ideal set(a). There are a few random locations where the value is not zero. When multiple copies of set (b) are available, the noise or randomness contributes everywhere. However the signal, the value "1" is contributed only at the cross. When the numbers are averaged, the noise can be either positive or negative and the sum will tend to be zero. However the signal is also summed and will remain. In order to show the similarity to the set (a) after the calculations, values shown in set (c) are normalized (divided by the number of trials).

In real life situations, sets like (b) will be observed and so a real signal that is actually present may be masked. By collecting sufficient amount of data and employing the averaging procedure, signal buried in noise may be recovered. In real experiments every measurement has an inbuilt averaging due to the physics of the system, in addition to any mathematical calculation that is employed. We will discuss the relevance of this in later chapters.

IV.5 Comparing uncertain numbers

Comparing two real numbers representing observations, each with an uncertainty is very difficult. At best both of these numbers can be the mean values of Gaussian distribution of trials. Consider two distributions with mean values 100 and 90 and equal standard deviation of 4. Now as mentioned earlier, 98.7% of the trials will ensure a value in the range 88-112 for the first ($100\pm3\times4$) and 78-102 for the second. There is thus an overlap. There is always a possibility that in one particular trial, the second number can be observed to be larger than the first. Uncertainty makes it difficult to compare such numbers. Thus ordering based on the ordinal value of numbers, which was the basis for mathematics, is disturbed by the uncertainty. While it is logical to state that $3 > 2$, a statement that $3(s) > 2(s)$ (where the (s) indicates the standard deviation in the value) is not. There is a finite possibility that there can be a trial in which the inequality is not valid.

The chi squared test described earlier permits one to make a statement "the probability that the observed data are caused by pure chance or randomness is x" or that "the data are statistically significant to 100x%". Since the probability is usually determined in the range 0-1, when a percentage is calculated it is multiplied by 100. The limitation of this statement in the context of how well we know was discussed in the last chapter. A similar method to compare two uncertain numbers is the "T-TEST". Given the two mean values, the corresponding standard deviations and number of observations (m1, m2, s1, s2, n1, n2), using the "T-TEST" one can determine a probability 'x' that the differences in the means are generated purely by chance. Equivalently, the difference between m1 and m2 is statistically significant to 100x%. As mentioned above, even if

m1>m2, in a given trial, the values obtained may not satisfy this inequality. Thus only a probability can be given. The use of "T-TEST" for justifying various "scientific" claims is so widespread that it is necessary to look at the details in some detail.

The first step of the "T-TEST" procedure is to calculate an "estimated standard error". The ratio of the difference in means to the "estimated standard error" is then determined and used to obtain a theoretical probability (as in the case of the chi squared test from a graph). The curve used for determining the probability from a "T-TEST" is shown in figure IV.3. The "degrees of freedom" is selected as (n1+n2)-2. There are special formulae to calculate the estimated standard error if s1,s2 are equal and when they are unequal (also when n1, n2 are equal and unequal).

If the observed difference between means is large compared to the expected standard error, it is probable that the difference in the means is significant. The difference between means is probably real. On the other hand if the observed difference between means is small compared to the expected standard error, the difference in the means could emerge from purely random causes and there is no significance to it. The use of the standard plot makes this qualitative statement quantitative. The value obtained from the graph is the probability that the differences between the two sets are caused by pure chance.

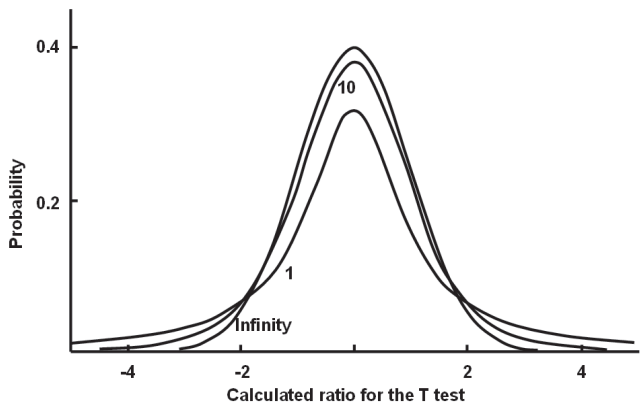

Figure IV.3 Plots used to calculate the probability that the difference in means is caused by random chance

As in the case of the chi squared test, usually a value of .05 is chosen as the acceptable boundary. If probability is lower, the differences are said to be statistically significant at 5%.

The probability distribution in the figure IV.3 looks like the Gaussian distribution with a mean of value zero and a standard deviation of one. The curve for infinite degrees of freedom is identical to the Gaussian. The picture shows that for smaller degrees of freedom or small sets of data the probability near the mean is smaller while that far removed from the mean (the tails) is higher. In finite sets of data, the probability of finding an "outlier" (a data point far removed from the mean) is higher than for a large set of data.

There are several problems with the "T-TEST" in the present context of trying to understand how well we know. The description of the "T-TEST" provided above is very basic. More complex mathematical and statistical discussions are available in the literature. However, the problems identified below as relevant for the present analysis are not addressed by more complex mathematics. Clear appreciation of these limitations as outlined in the next few pages will be most helpful in understanding how well we know in subsequent chapters.

The first limitation is that the probability obtained from a "T-TEST" depends on the number of data points available. To calculate the "estimated standard error" one divides the standard deviation with the square root of the number of data points. Thus as the number of data points is increased, the standard deviation does not decrease but the "estimated standard error" decreases. This reduces the probability of the data being the result of random causes, as determined from the plot. So with a sufficiently large number of data, any difference in mean can be made statistically significant to any desired accuracy.

To demonstrate this limitation, two sets of random numbers with a Gaussian probability were created as a test case each with 1000 numbers. The two sets had mean values of 10 and 10.5. The standard deviation was 1 for both. When "T-TEST" procedure was applied to small sets of 10 data points each taken from these large sets of 1000, the probability is .023. This is simply calculated in the

well known EXCEL spreadsheet software. If 50 data points each are selected from the same sets of 1000, the probability drops to .001. If the entire set of 1000 numbers are used the probability will drop to an astronomically low vale of 10^{-30}. Clearly if a limit of .02 was chosen, the means are not statistically significant when 10 data are available but statistically significant when 50 data have been taken.

Another issue concerns the use of the test with low degrees of freedom. A small set of data of say ten data points is available. The mean and standard deviation of nine of these values is determined. Obviously there are ten different ways of selecting nine data points each from the available ten data points. The first, the second or any other data point can be omitted to get a set of nine points. The mean values corresponding to each of these sets and the standard deviation are different despite only one data element being different between any two sets of nine. Table IV.2 shows the results. In the first row are the ten individual data points taken from the set of 1000 data with a mean of 10 and standard deviation 1. The second and third rows display the means and standard deviations obtained by selecting nine of the data points in the first row in turn. The third row gives the standard deviations for each set of nine data points.

Table IV.2

Calculations demonstrating the sensitivity of "T-TEST" probabilities to change of a single data point in small sets

9.35	8.99	9.99	11.1	8.77	11.4	9.66	8.48	9.42	9.49
9.68	9.70	9.74	9.63	9.51	9.76	9.47	9.67	9.80	9.69
1.00	0.99	0.97	0.99	0.85	0.94	0.76	1.00	0.90	1.00
0.063	0.067	0.076	0.049	0.018	0.014	0.012	0.059	0.027	0.065

Now we can compare each of these set of nine points with a single set of ten numbers drawn from the second set of 1000, those with a mean 10.5. The plan is to check if the difference between the means of the nine data points (mean10) and those of the ten data points (mean 10.5) is statistically significant. Since the means and standard deviations are changing, the corresponding "T-TEST" probability values will change as shown in the fourth row. Usually, a particular probability is selected as the limit for statistical significance. If as before a value of .05 is accepted, the decision regarding statistical significance is altered by a single data point. In some columns, the

"T-TEST" probability is higher than 0.05 and in some it is lower. For some of the sets the difference in means is significant at 5% for some of the others it is not. This difference is caused by a single data point that is changed in the sets. The conclusion of statistical significance is not robust for small number of data.

It is normally expected that the "T-TEST" can be used with small sets of data if the correct degrees of freedom are used. However, the above analysis shows that the conclusions drawn from small sets of data are not very reliable. Interestingly the range of the mean shown in the second row is smaller than that of the data in the first row. This is an example of the reduction of random errors by averaging as discussed earlier.

The use of a probability level for deciding on statistical significance can also be misleading. To understand this common mistake, consider the statement that the average height of men is larger than that of women with a statistical significance limit of 5%. This does not mean that the probability of finding a random woman to be taller than a random man is 5%. As shown above, the probability of significance varies with the number of samples used for calculating the "T-TEST" probability while the probability of a woman being taller must have a fixed or limiting value, for large amount of data. The test provides only a single point conclusion namely that the dif-

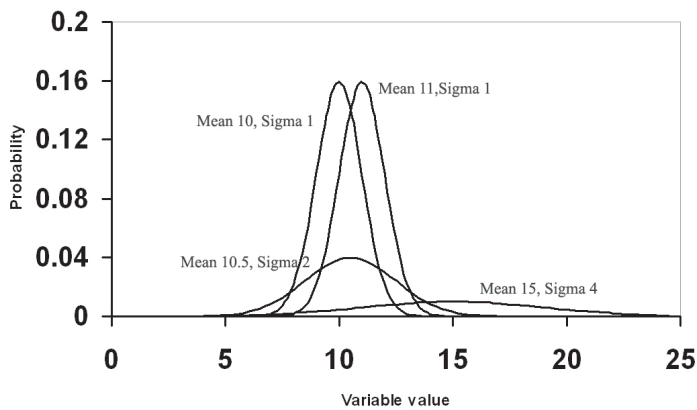

Figure IV.4 Overlapping Gaussian distributions with different values of mean and standard deviation

ferences are statistically significant. This can correspond to many different situations which are not distinguished by the test.

Consider figure IV.4 where several Gaussian distributions are plotted. The two large peaks have mean values of 10 and 11 and standard deviation 1 each. The smaller peak has a mean value of 10.5 and a standard deviation of 2. The fourth has a mean value of 15 and a standard deviation of 4. From a visual inspection, it may appear strange to state that difference between the peaks with mean values of 10.5 and 10 can be statistically significant. Most people without training in statistics would not consider these to be distinguishable peaks. They would expect that the term, difference in means are significant should be applied to two distinct peaks with very little overlap. But calculations show that with a hundred data points drawn from these two sets, the "T-TEST" probability is .05 and if 1000 data points are available the probability is as low as 10^{-10}. Thus if 100 data points each are available, the difference between the two means is significant to 5%.

The visual image emphasizes the significant overlap. Thus there is a probability that in a given trial the variable with a mean value of 11 can have values smaller than the variable with mean value 10. This leads to a question that in some sense is most important. Given that the difference between m1 and m2 in (m1, m2, s1, s2, n1, n2) is statistically significant according to the "T-TEST", what are the probabilities for observing m1>m2 and m1<m2. Such probabilities are most important for future discussion. Assume that the heights of men and women are collected in large numbers and shown to be Gaussian distributions about mean values. The difference in mean values is confirmed to be statistically significant. This means that the statement that the average man is taller than the average women is statistically significant. However there is a finite probability that during a random sampling, a woman can be taller than a man. This probability will be smaller than the probability that during random sampling the woman will be shorter than the man. However the magnitude of this probability will be important in assessing how well we know this and what use we make of this knowledge. To analyze this problem, a series of simulations have been performed using 10,000 random numbers with the required values of mean and standard deviation.

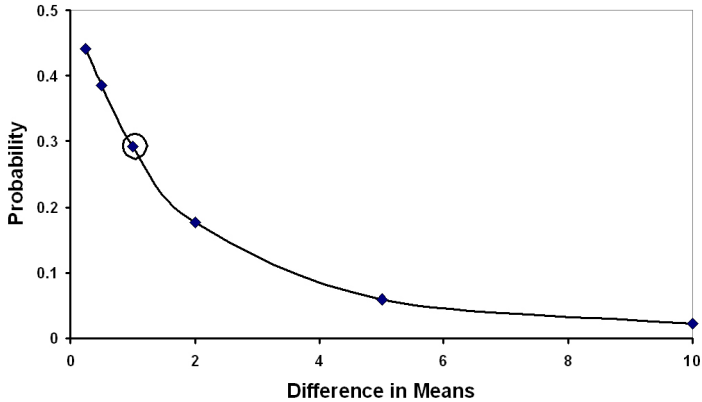

Figure IV.5 Slow decrease of the probability of a trial yielding a higher value from a set with lower mean value as the difference in mean is increased

The results are summarized in figures IV.5 and IV.6. Consider the two Gaussians discussed earlier with mean values 10 and 11 and equal standard deviation of 1. Random samples belonging to the two distributions are compared and the probability of drawing a number from the set with a mean vale of 11 that is smaller than a member from the set with a mean value of 10 is calculated. This is approximately 0.29 and the relevant point has been circled in figure IV.5.

Now the second peak is moved nearer or farther away from the other set with a mean of 10. In other words sets with mean values, 10.25, 10.5, 12, 15 and 20 each with a standard deviation of 1, have been investigated. The probabilities determined form the curve shown in figure IV.5. As the peaks move nearer, the probability can be expected to increase. If the peaks were identical, this probability has to be obviously 0.5 and the trend shows this would happen. As the peaks are moved apart, the probability decreases but rather slowly. Even when the difference in the means is 2, twice the standard deviation, the reverse probability is a significant 0.18, roughly 1/6. Thus if these peaks represented the heights of women and men, in about one case out of six, the random woman would be taller than the random man even if the differences in mean is twice the standard deviation. This probability does not depend on the statistical significance and is therefore more valuable than the “T-TEST” conclusion.

As was seen, when the standard deviation increases, the peaks become broader and the peak value decreases. To investigate the consequences for the probability, the two peaks were maintained at the respective mean values of 10 and 11 while the standard deviations were varied from 0.1 to 3. Once again the probability for a sample drawn from the set with a smaller mean being larger than a sample from the other distribution is determined. The data are plotted in figure IV.6 and the data for standard deviation 1 has been circled. As can be expected, the probability is very small when the peaks are very narrow. For example, when the standard deviation is 0.1, there is less than 2.5% probability of obtaining a value less than 10.8 for one peak and more than 10.2 for the other. As the standard deviation increases, the peaks become broad and overlap. One expects higher probability that a sample from the set with mean 11 will have a value less than sample from the set with mean 10. As the plot shows, the probability saturates at a value of 0.4. Calculating standard deviation is a meaningless exercise when the standard deviation is nearly equal to the mean but a standard deviation of 3 for a mean value of 10 is not in that category.

When the standard deviations are very large, the overlap between the curves is very large and the statistical significance cannot be easily evaluated. However the probability can show a significant bias. In this situation, the height distributions of men and women would be very broad but a random sampling would show that the

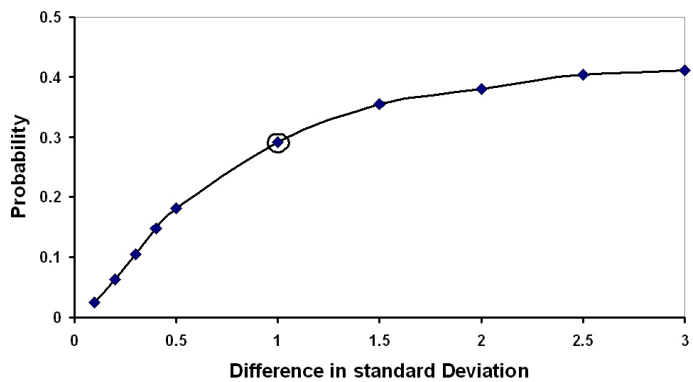

Figure IV.6. The increase in probability of the set with lower mean providing a larger sample during a trial as the standard deviation is increased

probability of a man being found taller could be significantly higher than 0.5! While "T-TEST" is not used very much in physics, it is very common in other branches of human endeavor. The above issues that can be easily observed with a little bit of mathematical calculation will be important for the discussion of how well we know in later chapters.

IV.6 Regression to the mean

As the concept of a mean, its relation to reduction of noise and probabilities has been explored, it is pertinent to discuss the issue of regression to the mean. Use and misuse of this concept contributes significantly to the problems being discussed. When the case of random events such as the toss of a coin was discussed, the series of heads and tails were presented. These showed significant clustering. There were series of consecutive heads and tails. The expectation that following a series of heads, the probability for a tail on the next try is high is called the gambler's fallacy.

This can be compared with the series of numbers with a mean of 10 and standard deviation 1 shown in figure IV.7. In this series of random trials, when a number far from the mean is observed the next element is usually closer to the mean. On the other hand when an observation is closer to the mean, the next element may also be closer to the mean. A set of three consecutive values close to one another

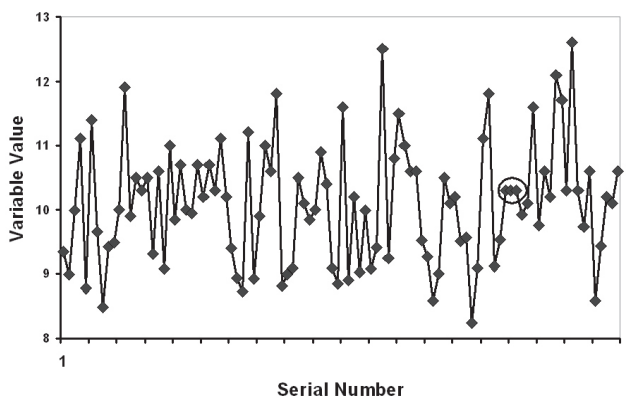

Figure IV.7 A series of 100 random trials for a variable with mean 10 and standard deviation 1, demonstrating regression to the mean

and the mean have been circled. This tendency of the sequence to revert to the mean value is called regression to the mean. Superficially, gambler's fallacy is similar to regression to the mean. When one encounters a man who is much taller than the average, his children are more likely to be shorter than the father and nearer to average in height. When a player in cricket or baseball fails after a series of superlative performances, this is attributed to the "law of averages", the more common name for the terms used in technical language, "regression to the mean" or "central limit theorem". Understanding the reason why the law of averages is valid in these situations but not when a gambler expects the next toss of a coin following a series of "tails" to be a "head" is most important.

A 10 sided die is rolled and the series of numbers plotted in figure IV.8 which can be compared to figure IV.7. There are significant differences. A mean as also a standard deviation can be calculated for this set of data. A value of mean close to 5 will be obtained since the numbers range from 1 to 10. But the mean is not the most probable value. The probability of a number 9, with a large deviation from the mean being followed by a number with a smaller deviation is not high. The probability for any number to follow the number 9 is the same, namely 10%. Regression to the mean is not observed in figure IV.8.

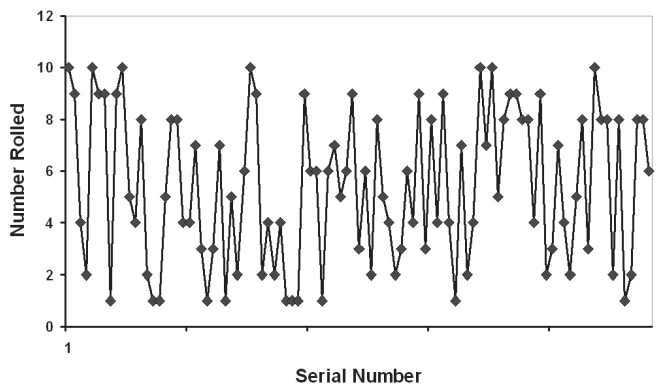

Figure IV.8 A series of 100 random trials of a 10 sided die demonstrating lack of regression to the mean

It must be noted that the series with average 10 and standard deviation 1 was also claimed to be a random sequence. However the Gaussian distribution of probability ensures that there is an underlying or "hidden" linkage between successive numbers. Rather than use a single die with large number of faces, the numbers from two or more dice can be added to get numbers in a range. For example, a number ranging from 1-24 can be obtained by using 4 dice each of which give a number ranging from 1-6. When these were discussed earlier, it was pointed out that one observes a Gaussian distribution of numbers when random numbers are added. In this case the hidden aspect is the number of permutations and combinations that can give a sum. If four dice are used, there are many more combinations that can lead to a value of 12 rather than 24. Thus a number 12 is more probable than 24. If four dies are employed, one throw resulted in a value of 24, it can be expected that the next throw would result in a value closer to the mean value of 12. If only one die with 24 faces is used, the probabilities for 12 and 24 are both the same. Thus the most important requirement is for the hidden reason responsible for a Gaussian distribution. Once there is a Gaussian distribution, there will also be a regression to the mean.

Regression to the mean becomes relevant in real life situations when the effect of a procedure on the system is investigated. Thus instead of comparing two variables the system is tested before and after an attempt is made to enhance or decrease the mean value. For example the effect of diet, exercise or training on the performance in tests or competitions is often evaluated in this way. Even if the intervention (diet or exercise) has no influence there can be a spurious change if the first sample is not representative (it is far from the mean). We shall discuss several issues of regression to the mean in later chapters.

IV.7 When the distribution is not Gaussian

The ideal experimental situation would be when the experimental quantities are represented by Gaussian peaks. If the mean and standard deviation are also well defined, it means that further investigation would be useful. If we measure the weight of an object

in a large number of instruments or do it a number of times, we get such a peak. There are variations but these are small and random. This ensures that we can compare two weights and we can study the influence of other parameters.

However, many sequences obtained in real life do not result in such distributions. There could be two reasons for this. One important possibility is biased sampling. The underlying distribution may be Gaussian but the process of selecting the finite number of data points for analysis may not be properly representative. An example is shown schematically in figure IV.9. In the first part, the dots are randomly situated and the different circles of same area contain roughly same number of dots. In the second part, even though the dots are still the same, instead of circles odd shapes have been chosen for sampling. The number of dots in each is not even approximately equal. Such a situation is called biased sampling and the problem can be identified only as a case by case basis (as with other similar problems encountered before).

On the other hand, the distributions may not be Gaussian at all. For example if we determine not the weight of the individuals but the wealth of every individual in the world, the result would not show a Gaussian distribution at all. One consequence of non-Gaussian distributions is that the mean median and mode are not equal. Several situations of non-Gaussian distributions are shown in Figure IV.10. Even when the distribution is symmetric and the mean median and

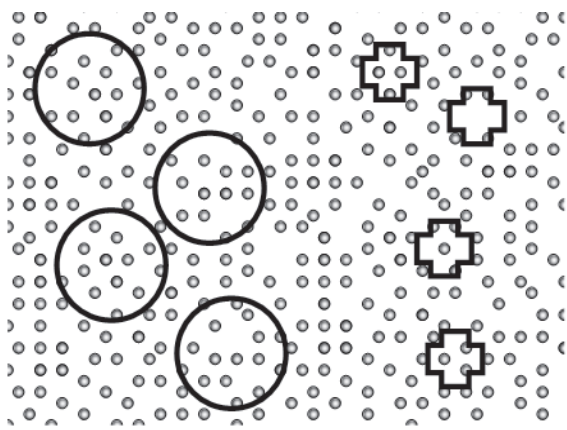

Figure IV.9 Pictorial description of fair and biased sampling

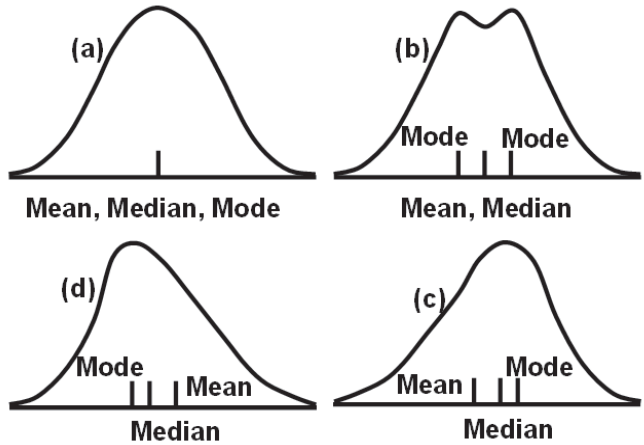

Figure IV.10 Gaussian and non-Gaussian probability distributions

mode are all equal, the curve may not be necessarily Gaussian. For example, if the distribution is a semicircle, median mode and mean are equal but the curve is not a Gaussian. In all these cases, the standard deviation can still be calculated. This is simply obtained by determining the mean and the deviations of each trial result from the mean, summing the squares of the deviations and finally determining the average. However, the probabilities associated with the Gaussian are not valid. In all such cases, interpreting the statistical quantities in real life is extremely subjective. So comparing two mean values obtained from such distributions is extremely tricky. The probabilities for the m1>m2 and m1<m2 described earlier are not applicable and the expectation values that can be defined in the Gaussian context do not apply. The regression to the mean will not be as clearly applicable. When practical problems are discussed in later chapters these limitations will all be highlighted.

The analysis of the present chapter shows clearly some of the inherent limitations in the effort to quantify knowledge and thereby have a realistic and honest assessment of how well things can be known. Even if numbers are associated with real life phenomena, comparing two such numbers is inherently difficult. As mentioned before, even making the simple assessment that one number is larger than the other can only be made provisionally and approximately just as in the case of our effort to identify a fair coin in the earlier chapter.

V
HOW TO RELATE UNCERTAIN NUMBERS

V.1 Relationships between numbers

The simplest way to look at a relationship between two numbers is to recollect the definition of π, the ratio of the circumference to the diameter of a circle. Obviously as the diameter of the circle is changed so does its circumference. We have already seen that this ratio is a universal constant known to a very high accuracy. However if we assume that this is not known, one can determine the circumference of a number of circles and their diameters and try to understand the relationship between the two numbers. This was probably how the ancient cultures came up with the initial values for π. As discussed in the previous chapter, numbers used to quantify things of interest have uncertainties or errors associated with them. Thus the key to understanding how well we know anything is to understand how to relate uncertain numbers. When a large collection of diameters and the corresponding circumferences are available, they form sets (d, c).

The simplest way to organize these pairs is to arrange them in ascending order of the diameter (d). Then a simple examination of

the pairs confirms that the circumference values are automatically in the ascending order. An example is {(1,3.2),(1.3, 4), (1.6,5.1), (2,6.3)}. In each pair of numbers, the first is the diameter and the second is the circumference. If we divide the second number by the first, it would be apparent that the ratio is not the exact value of π as defined earlier. Thus there is an error that is the result of practically determining the diameters and circumferences. In this example, there is no case of an observed decrease in the circumference when the diameter is increased. However, if we change the diameters by small values, it is possible to observe the following set. {(1,3.2),(1.03,3.19),(1.06,3.39),(1.09, 1.38)}. When the increments in diameter are small, due to errors, it is difficult to confirm that as the first variable increases the second also does increase. As in the earlier discussion of the mean of a number of observations and later while comparing uncertain numbers, uncertainty complicates matters. Assume that in addition to the above (d,c) there is another pairing (d, w). For example if we restrict ourselves to circular disks, w could be the weight. First assume that all the disks are made of the same material. Assume further that a perfect ranking order is found in both (d,c) as well as (d,w). This means whenever there is a larger value of d, larger values of both c and w are observed. Thus d is related to both c and w. However, the two the relationships are different. We know that the relationship between the diameter and circumference is absolutely true. It is only measurement that causes the errors mentioned above. If we consider the weights, it is quite easy to realize that the weight of a larger wheel made of a light material like wood could be smaller than that of a smaller wheel made of iron. The first relation does not alter with the thickness of the disk, the second does. In such situations, assigning a quantitative descriptor to the relationship is desirable. This would enable more precise comparison of various relationships. This is similar to the problem earlier associated with the mode. Simple ordinal description of being larger or smaller is not sufficient.

V.2 Correlation coefficients

Taking the entire set of data, the ranking of each of the variables can be determined. For example, the series {(1.0,3.2), (1.03,3.19), (1.04,3.25), (1.06,3.39), (1.09,3.38), (1.1,3.4)} is converted into a series of ranks {(1,2), (2,1), (3,3), (4,5), (5,4), (6,6)}. The two

variables, diameter and circumference are individually ranked. Since 1.0 is the lowest value of the diameter it has a rank 1. Since 1.1 is the highest value among the six available values of diameter it has rank 6. The circumference associated with the first pair, 3.2 is larger than that in the second pair namely 3.19. Thus it has a rank 2. Thus after ranking, the first pair is (1,2). There sets like (6,6) and (3,3) which are perfect. The presence of a large number of such cases could be taken as a guide to a qualitatively better relationship. One can also find the difference between these ranks leading to a series(-1,1,0,-1,1,0). Small values are another indication of a "good" relationship.

It is possible to quantify the differences more precisely. One way is to determine an average for original data. Then a large deviation from the mean value of one variable would be expected to result in a large deviation from the mean value to the related variable. Thus if a mean value for the diameter of the circle is first determined and the mean value of circumference, a sample circle which has a diameter far away from the mean value must have a circumference value far different from the mean circumference too. So one can have sets of data defining the deviations from the mean. For the given data, the means are 1.053333333 and 3.301666667 and sets (-0.05333,-0.101667), (-0.02333,-0.111667), (-0.01333,-0.051667), (0.00667,0.088333), (0.03667,0.078333) and (0.04667,0.098333). (While these numbers have been given to such a large number of digits, these are actually not correct. If the diameter is measured to an accuracy of the second decimal place, the average cannot be known to the ninth decimal place. This however is not being discussed since such rules are specific to a given problem.) Thus one can confirm that a larger deviation of the circumference from its mean is always found when the deviation of the diameter from its mean is large. When one is negative the other is also negative. When one is positive the other is also positive. All this descriptive effort only provides marginal improvement of our knowledge and still cannot help in comparing different relationships.

When a large amount of data is available, a numerical correlation coefficient can be calculated. It is a value ranging from +1 to -1. A correlation coefficient of +1 indicates perfect positive correlation, when one variable increases the other also increases

proportionately. When it is -1, the correlation is still perfect but opposite, when one increases the other decreases proportionately. If the two variables are not related one obtains a value of zero for the correlation coefficient. Thus once can asses the relative strengths of two correlations by comparing their correlation coefficients.

The deviations from the mean described above can be employed to calculate the coefficient. The deviations in two variables being considered are multiplied and summed for the entire set. This is divided by the means square deviation, sum of the squares of the individual deviations for each of the two variables. This ensures that it all gets normalized to a value in the range -1 to +1. This is called the "Pearson's correlation coefficient". As with the chi square test, squaring emphasizes a large deviation and also ensures that positive and negative deviations are treated identically. A correlation coefficient can also be calculated using the ranks. It is called "Spearman's rank correlation coefficient". Here the average rank is determined but the rest of the mathematics is identical to the Pearson coefficient.

These coefficients can answer the question raised earlier. Comparing the relative strengths of two relationships, if the correlation coefficient for (d,c) is closer to 1 as compared to (d,w) it would be a reasonable conclusion to say that the first of the relationships is stronger than the other.

V.3 Limitations of correlation coefficients

Calculating the Pearson or Spearman correlation coefficients is extremely simple and with the easy availability of computer software, effortless. This encourages extensive use of the procedures often without understanding the limitations of the procedure. Major limitations are discussed below. The first limitation of the correlation coefficients is that their values are smaller when the ranges of the two variables are restricted. Using the data of diameter and circumference presented one obtains a small correlation coefficient (0.3) since a very small range (10%) is employed. In figure V.1, when the range of X is restricted to the interval (0,1) the correlation coefficient is approximately 0.6, much smaller than the values of nearly 0.79 obtained when the range of X is larger.

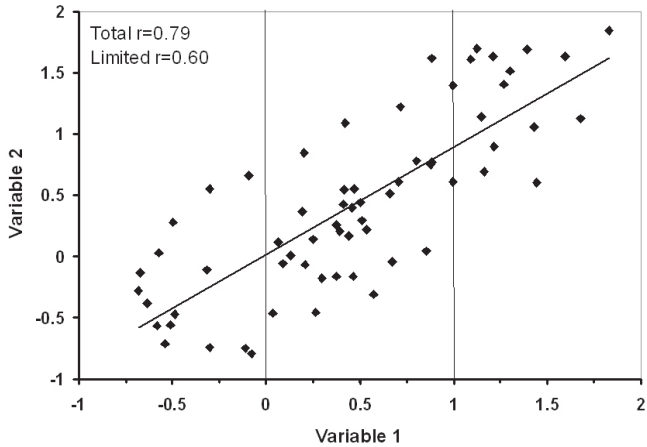

Figure V.1 Reduction of the correlation coefficient when range of variables is reduced

In addition to this, as in the case of the arithmetic mean, the presence of a few data points far away from the mean influences the value of the Pearson correlation coefficient much more than they influence the ranking coefficient. However, a value of the ranking coefficient close to 1 does not ensure that the change in one variable is proportional to the change in the other. The more commonly used Pearson coefficient assumes that the deviation in the second variable is proportional to the deviation in the first. This is therefore called linear regression. An ideally correlated data will fall on a straight line which is mathematically described by the equation $y = mx + c$ where m, the slope of the line and c the intercept are constants.

The assumption of a linear relationship causes its own problems. Figure V.2 plots four famous sets artificially created by Anscombe. All sets have identical values of "Mean of x (9.0), Variance of x (11.0), Mean of y (7.5) Variance of y (4.12), Correlation coefficient (0.816) for a linear regression line ($y = 3 + 0.5x$)". The first one (top left) seems to be distributed normally and corresponds to what one would expect when considering two correlated variables. The second one (top right) has an obvious, visible non linear relationship between the two variables and the correlation coefficient is really not relevant. In the third case (bottom left), the distribution is linear, but one point lies outside the general trend. It is visibly appar-

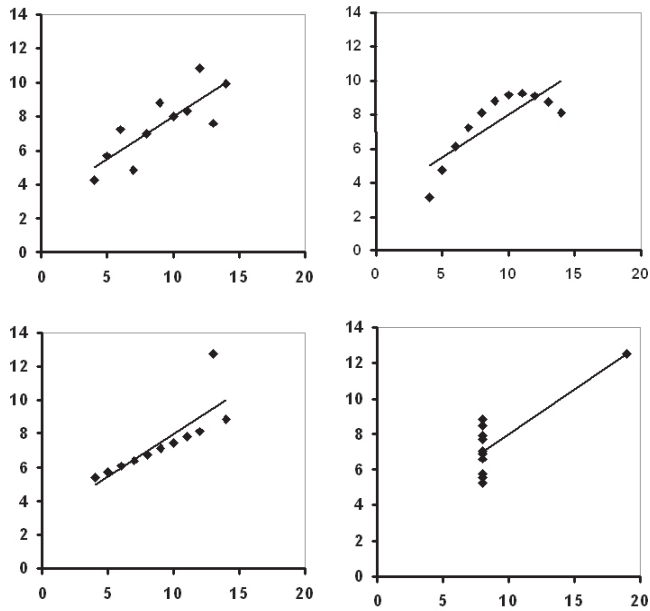

Figure V.2 Anscombe's famous picture showing the various possible conditions for high correlation coefficients

ent that a different linear relationship with a smaller slope, passing nearer to the rest of the points is more relevant. Finally, in the fourth example, one outlier produces a high correlation coefficient, even though the relationship is not linear. These examples have been designed to specifically highlight the limitations of the correlation coefficient.

Just as a high value of the correlation coefficient can arise from many causes, a zero value of the correlation coefficient cannot ensure that the two variables are independent of each other. Figure V.3 shows a graphical representation of large sets of data. Each dot in the picture represents one pair of values (x,y). In each case the corre-

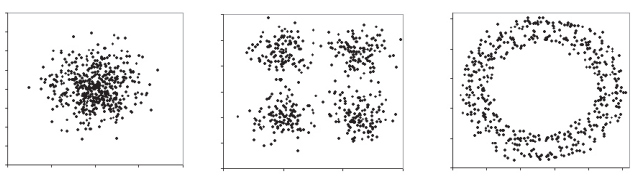

Figure V.3. Possible scenarios for zero correlation coefficient

lation coefficient can be calculated and will be found to be fairly close to zero. The first example is what one expects to see when there is no clear linear relationship. This image corresponds to the correct case of the coefficient being zero. There is no change in one of the variables even when the second varies.

In the second and third pictures, there is significant order in the picture but the numerical value of the correlation coefficient is zero. In general there can be two reasons for this kind of observation. One possibility is that the functional dependence is not linear. This is similar to the second example in the Anscombe data. In addition to a nonlinear relation, if the first variable is symmetrically distributed about 0, the correlation coefficient would also be zero. For example if the second variable is the square of the first, they are mathematically related but the calculated value of the correlation coefficient will be zero. A second possibility is that the variables change with time in a prescribed way. The data in the second and third pictures were obtained by artificially modifying the data in the first pattern. In the first case equal parts of data sets (x,y) were added to (0,0), (0.25,0)(0.25,0.25) and (0,0.25). In the second the x and y variables were converted to trigonometric sine and cosine functions. These are specified mathematical functions but the correlation coefficient fails to detect this order.

V.4 Distribution of variables

In an earlier discussion it was seen that the mean and the standard deviation contain more information if the variable follows a Gaussian distribution. While calculating the linear correlation coefficients, it is not considered necessary that the variables are normally (Gaussian) distributed about their respective mean value. If the data are Gaussian, most of the samples will emerge from the central region. By definition of the Gaussian peak, 65% of the samples will fall within mean ± standard deviation. Thus there are few data points in the rest of the range. Even if there is no true linear dependence, a good correlation coefficient may be obtained because of the limited range of data. For example, if the positive outliers have a negative deviation and negative outliers have a positive deviation, which could be caused by non linear relationships, the correlation coefficient will

increase if the range is decreased. This is the reverse of the example cited earlier where reducing the range of the variables decreased the correlation coefficient.

Thus a uniform distribution in the range of interest is more suitable for using the linear correlation approach for comparing the strengths of various relationships. Consider for example, the relationship between IQ numbers and earning capacity. This may show a linear relation. However, it is entirely possible that extraordinary high values of IQ would result in more than a linear increase in earning capacity and extraordinary low values decrease the earning capacity far below the linear relationship. When sampled, most common examples being in the mid range, a linear relationship can be observed, the actual relation might however be highly non linear. The key issue that is often ignored is the range over which a linear relationship is actually obeyed.

V.5 Dependent and independent variables

The above description treats both the variables as equivalent. In the data sets of diameter and circumference, there is no distinction between the two. In general, when data is collected by observation, this is the correct position. However when there is the possibility of experimentation, one of the variables becomes the independent variable and the second the dependent variable. In the first example of the diameter and circumference of a circle, instead of finding wheels to measure, a flexible spherical balloon can be chosen. It can be blown up to any desired diameter and the corresponding circumference measured. One can obtain in principle, a series of data points. Also by controlling the pressure one can repeat the measurement for a given value of the diameter. Thus the diameter can become the independent variable. Corresponding to a given value of diameter one gets a number of values of the circumference. This immediately points out the two important attributes of the independent variable. It is possible to repeat the measurement for a given value. It is also assumed that there is no error in the variable. The error in the observed values of the circumference for example would include a part due to the diameter being recorded incorrectly. However, this is ignored and the entire error attributed to the dependent variable. A further feature is the

assumption that the value of the independent variable can be set to any desired value. Thus if two measurements are taken with diameter values of 1.0 and 2.0, it is assumed that measurements can be repeated for values of 1.5, 1.75 etc. Obviously, it is expected that the values of the dependent variable obtained by repeated trials would constitute a Gaussian or normal distribution with a well defined mean and standard deviation. Thus error in independent variable is assumed to be zero and that the error in the dependent variable is due to multiple random sources. As discussed earlier the summing of these random errors results in a Gaussian distribution. In figure V.4, the errors in the dependent variable are indicated by error bars that represent range of possible errors. It could for example be three times the standard deviation giving a 98% confidence in the value. The error bar places one immediate restriction on the possible correlations that can be inferred from the data. In the picture, a straight line has been indicated to represent the correlation. Unlike the earlier case, if the line does not fall within the error bars, validity of the relation represented by the line is questionable. It is also possible to draw many straight lines and other curves for example a parabola passing through all the error bars on the data.

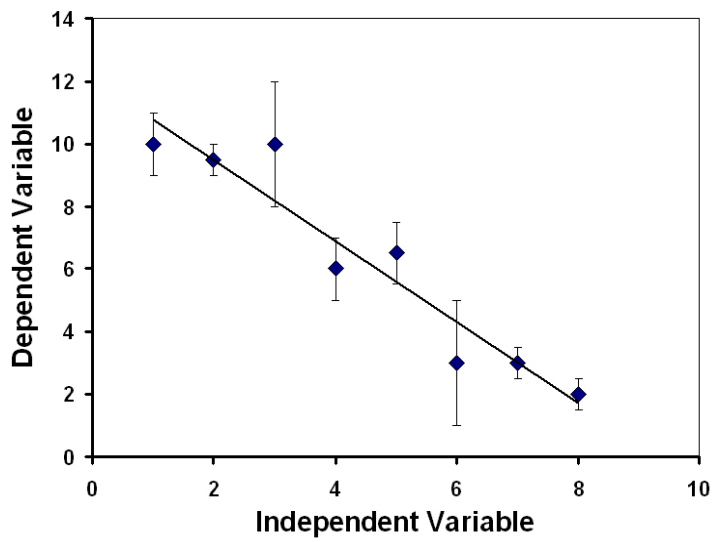

Figure V.4 A representative example showing dependent and independent variables with the error bars

V.6 Linear least square fit and Regression

What if a number of straight lines or even different curves pass through the various error bars shown in figure V.4? Selecting one requires a proper procedure with a logical justification. A justified functional dependence is obtained by a process called a least square fit. We first limit our discussion to straight lines. We assume that a straight line has been drawn. The data points represent given values of the independent variable and the corresponding dependent variable. In view of the errors, the line will not always pass exactly through the data points but will lie either above or below them as is the case with the data in figure V.4. The line should however pass through the error bars drawn. The difference between the data point and the point at which the line crosses the error bar is called the deviation. These deviations are as usual squared and summed. It is obvious that if a large number of straight lines are drawn, the value of the sum of the squares of deviations will be different for different lines and that there will be one line with a minimum value. This line is called the least square fit.

It is at once obvious that the linear least square fit must be related to the Pearson regression coefficient which also reflects a linear relation between the variables. As already mentioned earlier, a linear relationship between the two variables x and y is given mathematically by the equation y = mx + c. Once a least square fit is performed values of m and c are obtained. Now consider as a hypothetical case,

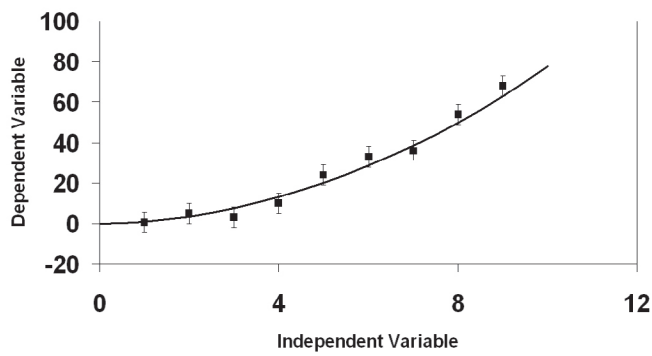

Figure V.5, The schematic of a least square fit of deviations

y to be the independent variable. Thus we can for the same data obtain another least square fit x = m'y + c' with two different constants m' and c'. The regression procedure, as has been mentioned earlier, treats both variables as equally prone to error. The Pearson coefficient is equal to $\sqrt{(mm')}$. This clearly reflects the essential symmetry between the two variables that is implied in the regression analysis. This equivalence between the linear least square fit and the Pearson regression is not valid in the case the relationship between the variables takes functional forms other than the straight line. In figure V.5, the least square fit has been performed with an equation $y = mx^2 + nx + k$, or a second degree polynomial as it is technically called. In this case the error bars are all considered to be equal. This is more common than errors being different for each data point as in the example shown in figure V.4. Obviously $x = m'y^2 + n'y + kc'$ is not a proper function to describe the relationship. So a regression analysis is not a meaningful way to relate data if the relationship has a mathematical form other than that of a straight line.

The least square procedure however can be applied to any mathematical curve, not merely the straight line. It is in general not necessary to try out a large number of lines or curves and select the one with minimum deviation. Identifying the "least" square fit is possible for many mathematical curves including the straight line. The procedure depends on the ability to differentiate such curves as discussed below. All this however does not answer the question posed at the start of this section. What if several different curves pass though the error bars and are thus reasonable possibilities? Even with a least square procedure the only limit is that the sum of the squares of the deviations is smallest for the particular function and smaller than the maximum error expected in the data. One can obtain least square fits for specific straight line, parabola or any other mathematical function satisfying the above criterion for the same set of data. One of these cannot be selected from this data alone.

V.7 Functional dependence

Understanding the concept of differentiation and integration, mathematical procedures used in least square fitting is also very relevant for some of the discussion in later chapters. As a consequence

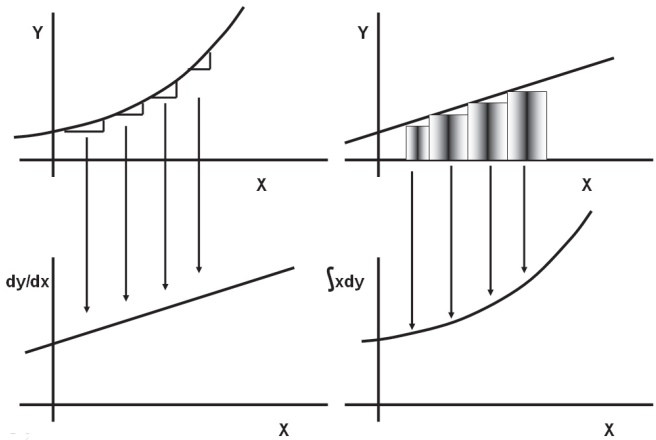

Figure V.6 A schematic explanation of differentiation and integration

a small digression to describe these is necessary. The designation of one of the variables as "independent" is not merely a matter of nomenclature. As mentioned earlier, this implies that its value can be altered continuously. Thus the independent variable acquires the mathematical characteristic of a real number where, as with rational numbers, between every two real numbers however close, another is always present. The dependent variable also is a continuous variable and the values obtained are a sample from the Gaussian distribution as mentioned above. The dependent variable being continuous permits the use of more advanced mathematical procedures most importantly differentiation and integration. Without going into the mathematical details, it is sufficient to recognize that differentiation results in a value that is equal to the slope or gradient of the dependent variable. Integration results in a value equal to the area under the curve as shown in figure V.6. The primary requirement for these mathematical procedures to be employed is that the curve itself does not have any discontinuities or cusps such as shown in figure V.7. If it is assumed that in the range of interest of the independent variable, there is no discontinuity and the dependent variable exhibits a smooth variation, it is possible to describe the relationship between the two as a mathematical function. When the straight line is represented as $y = mx + c$, the variables, x and y are real numbers as are the constants m and c. Such an equation is a called a simple, functional dependence between two variables. Functions could be algebraic meaning only

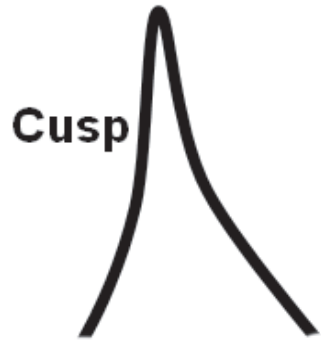

Figure V.7, A cusp or discontinuity

arithmetic operations such as addition, multiplication etc are employed. There could be more complex mathematical functions from advanced mathematics. The curves may look simple like a parabola or an exponential curve. They may be complex spirals or repetitive like a wave. In addition to the dependent and independent variables (x and y in the present description), the mathematical functions will have some constants. In the case of the straight line, m the slope and c the intercept are constants. Different lines are obtained by changing these values.

Given a set of data, consisting of a series of values of an independent variable and the corresponding dependent variable values, we can superimpose a plot on the data and alter the constants so that one line or curve with a minimum value of the sum of the squares of deviations. The least square fit procedure can provide the best curve of a particular type for example all possible straight lines. Each of them will have a net value of the sum of squares of deviations and the line with the lowest deviation can be identified as the best possible fit to the data

The identification of a unique or best fit is mathematically possible only for a class of mathematical functions which are termed linear. This should not be confused with the linear meaning a straight line in other contexts. After a number of differentiations have been performed on these functions, one obtains a constant value that does not depend on the independent variable. Then the problem can be solved and one unique set of values of the constants with minimum

sum of squares of deviations can be identified. Then depending on the error bar a few lines close to the minimum can be identified as being equally probable.

However even for a simple well established function such as the power law ($y = mx^a + c$ with "a" not being an integer), this procedure does not work. In the present discussion, these functions where a least square fit cannot be determined are nonlinear. In such cases, one has to follow a try and repeat exercise for various values of "a, m and c" to compare the sum of squares of deviations. With modern computers it is not a problem to obtain a curve with a low enough least square deviation. However there is no logical reason to believe that this is the lowest. The possibility of other curves with different values of the constants but equally low or even lower value of the sum of squares of deviations exists. Thus there is in principle an uncertainty in identifying the curve as the lowest possible.

In the discussion above we have considered two functions specifically. The power law curve $y = mx^a + c$ and the straight line $y = mx + c$. It is completely meaningless to compare the sum of the squares of deviations for these two curves for a given set of data. If by a happy coincidence the sum of squares of deviations for one of these is much larger than the error in the data while it is much smaller for the other, selection is possible. Ideally, the choice of the function, the power law over the linear relation for example must be based on grounds other than the comparison of error values and least square deviations. The same is true if one is comparing two linear functions for example the straight line and a parabola. At the very least, post facto justification of the reason for the choice has to be based on arguments other than the data. Choosing the suitable functional forms is the basic essence of how well we know and will be extensively discussed in the next part.

V.8 Extrapolation and interpolation

Even though the process of deducing a functional dependence linking the available data has many possible pitfalls, this also has a unique capability. The procedure enables one to predict the outcome of an experiment. If there is no prediction there is no science. If the

data consists of a sample of independent and corresponding dependent variables, the identification of a functional form enables predicting the value of the dependent variable for a new value of the independent variable. Going back to the first example of diameters and circumferences of circles, if the data consists of circles of diameters 1,2,3 … units and the corresponding circumferences, the circumference for a circles of diameter 1.5, 2.5,3.5…. can all be determined, even in the absence of experimental data. The predicted values can be then verified practically to increase our confidence in the functional form employed. Practical verification can be performed for a finite number of values but the procedure permits one to predict the results for an infinite number of possible values. This is the basic advantage of the process and is at the core of the effort to answer how well we know anything. If the independent variable has values in the range x1 to x2, the prediction can be performed both when the new value of x is in the range x1 to x2 in which case the process is said to be interpolation or when it is either less than x1 or greater than x2 in which case the process is called extrapolation.

From a philosophical point of view neither is justified. Philosophically this is even worse than the problem of induction that was mentioned in the first chapter. That referred to the repetition of the same thing as in "how does the fact that the sunrise was observed a million times imply that there will be one tomorrow?". Here one is making some observations and predicting what has not been observed so far at all. Surely a great problem! For the present this will be disregarded as with all philosophical conundrums.

Interpolation and extrapolation result in more practical problems. Except in the rare situation of deterministic chaos mathematically calculating the interpolated or extrapolate values is very simple. Interpolation and extrapolation are mathematically justified. The additional numbers are already contained in the brief functional form. To come back to the first example of the chapter, in mathematical form $c = \pi \times d$. This formula already contains in itself the numbers corresponding to c for every possible value of d. However when a functional form is deduced from experimental data, the relationship is not absolute. There is always the possibility that the real experimental data may throw up surprises. In the next chapter, many excit-

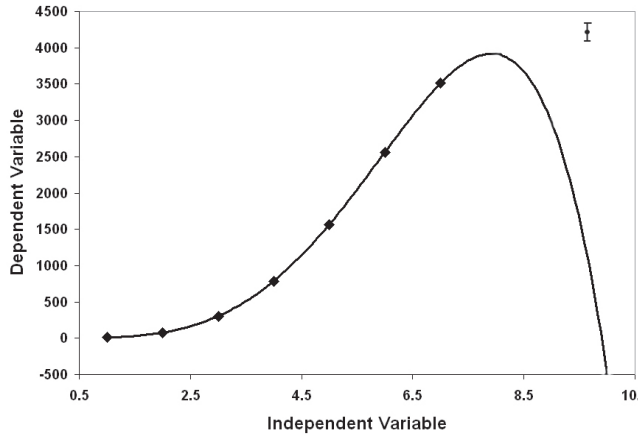

Figure V.8 An unwarranted mathematical least square fit

ing examples will be discussed. Here an amusing example is presented. Consider the data shown in figure V.8. The population of a city or even the world would show similar increasing tendency. It is possible to perform a least square fit of the data using many mathematical functions. Figure V.8 shows one curve which is obviously acceptable within the range of data available. The sum of the squares of the deviations is smaller than the errors. Thus one would use such a fit to extrapolate and find the value of the dependent variable where experimental data are not available. If this curve is employed to extrapolate beyond the range of the data, it can show a value zero or even negative. This is completely contrary to the trend reflected in the data itself. If such a function was used to fit population data, one gets an absurd or amusing prediction that the population will be negative soon! This is a classic example of misuse of the process of extrapolation. It is important however to note that the absurdity can only be identified post facto. Only when the particular extrapolation has been attempted is there reason to realize the absurdity of the prediction. This serves as a warning of the pitfalls that are present in this enterprise of drawing scientific conclusions based on experience.

———————

Summary

What Can Be Learnt From Mathematics

The famous quotation from Lord Kelvin advocates use of "numbers" to improve knowledge. The description in the previous four chapters of the basic mathematical tools that are available highlights the risks involved in using numbers and mathematical tricks. If care is not taken the process of using mathematics can degenerate into GIGO (Garbage In and Garbage Out). The whole purpose of the present effort is to develop the capability to identify such situations. The overview of mathematics shows that the simplest of mathematical operations, addition implies the existence of infinity. Repetitive use of numerical operations can provide not only order in the form of frequency doubling but also chaos. On the other hand there is a lot of "apparent" order in randomness as seen in the presence of ordered sequences among the random result of tossing a coin. The best of mathematics cannot be applied from one area to a closely related one while some things such as the randomness of the sequence of tosses can only be known approximately. The accuracy with which something can be known depends on the base line value. When something is repeated, and there are errors. Trying to take an "average" is not a trivial problem. Comparing two approximately known numbers (all experimental data are approximate) is a very complicated

issue and making even a simple relational statement that one is larger than the other turns out to be very tricky. Finally the ability to use mathematical functional dependencies for interpolation and extrapolation involves many exciting risks. Given this catalogue of the limitations of mathematical techniques the quotation from Lord Kelvin may appear strange. However the true appreciation of the scientific endeavor requires one to know the stupendous success that has been achieved with this imperfect tool. Most important for the goal of answering the question "how well do we know it" is to recognize the limitations which will highlight the accomplishments.

Part Two

Learning From Physical Sciences

There are several reasons for discussing at some length, "how well do we know physical science". The first and foremost reason is the exemplary accuracy with which we do know things. For example, in quantum physics, the value of the magnetic moment of an electron in some normalized units is measured experimentally as 1.00115965221 and calculated theoretically to be 1.00115965246. For the present discussion it does not matter what a magnetic moment is. If the diameter of the earth were determined to this accuracy it would be exact to the thickness of a human hair.

The second reason for discussing physical sciences is the accepted scientific "wisdom" that all chemistry is physics and all biology is chemistry which means that it is also physics. Physics is seen as a reductionist methodology and there is severe resistance to this idea not merely on the part of philosophers who term this "scientism" but even on the part of many scientists. It is of course a pity that all such discussions fail to basically understand the way physics actually is (in contrast to the philosopher's idea of what it is). Just to put matters in perspective, all physicists know that the "three body problem" cannot be exactly solved in Newtonian Gravitation,

the two body problem cannot be solved in General Relativity, the self energy problem has not been solved for one point charge in Classical Electromagnetic Theory and the Vacuum Ground State cannot be solved exactly in Quantum Mechanics. So if one is after exact solutions, no body (Vacuum) is already too many. The beauty of physics is how much can be known despite this.

Another reason for the discussion in the next few chapters is to bring out the actual interaction between experiment, theory and prediction so that the terms like reductionism, holism, contingency that are employed in philosophical discussion of physics are properly understood in the physics context. This will delineate a proper methodology for evaluating more complex issues subsequently.

VI
HOW MEASUREMENTS ARE RELATED IN PHYSICS

VI.1 Functional dependence in physics

Physics began with measurement of quantities of everyday experience. One of the oldest and certainly the most famous is Galileo's measurement of the period of oscillation of the candelabra in the church by comparing it with his own pulse rate. This tendency to measure one observable quantity in units of another physical quantity is central to physics. In this case the period of oscillation of the candelabra is compared with the period of the pulse. Thus quantities measured in physics are numbers usually associated with "units".

Galileo noted that the time taken for the swing of the candelabra was the same irrespective of the size of the arc through which the candelabra was swinging. Following this initial observation Galileo developed the law of simple pendulums. The period of oscillation of a simple pendulum, at the simplest a string to which a small stone is attached, depends only on its length. This experiment is verified by every high school student of physics, using a stop watch to measure the time period in seconds which becomes the "unit" of measurement.

Just as Galileo sought to establish a relationship between the length of the pendulum and its period, every physics investigation seeks to relate one experimental parameter with another. In the language developed in the last chapter, one of the variables is always the independent variable and the other is the dependent variable. In Galileo's experiment, the length is the independent variable. There are many different functional forms that are observed in physics. A few of the simplest mathematical relationships will be discussed next.

VI.2 The exponential dependence

The exponential relationship between the dependent and independent variables is certainly very common and important in physics. Consider the intensity of light at different distances from the source. This would of course depend on whether one is referring to light propagating in air or water. It will also depend on whether there is a fog or if the water is murky. But above all else it will depend on the amount of light that is available. Consider a possible situation where half the intensity of the light available on the surface of water remains at a depth of one meter. What amount of light is left at a depth of two meters? The question is logically answered by pointing out that since half the intensity of light at the surface is available at one meter depth half of that or one fourth the intensity would have

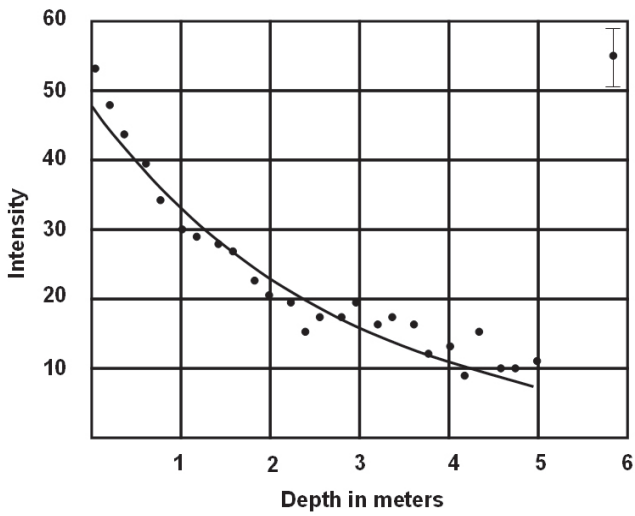

Figure VI.1 Variation of light intensity with depth of water

reached a depth of two meters. The differences between air, fog and water will only be the depth at which the intensity has reduced to half the value at the surface.

This type of dependence where the change in a given quantity is dependent on the quantity itself is called an exponential change. The typical example is shown in figure VI.1. The dots are the data and the error bar is shown in the top right corner. The exponential curve does not pass though all the data but the deviations of the data from the curve are smaller than the error and so the functional dependence is acceptable. This is an exponential decrease in the intensity of light with the depth of water. Similar dependence of many other physical parameters are known. In addition to decreasing levels of light or sound with distance, variation of voltage on a discharging capacitor, change in temperature due to cooling, speed of chemical reactions and decay of radioactive atoms display this behavior. Variables can also increase exponentially. If one ignores complexities like the carrying capacity of the earth and the finite lifetime of any living being, population would increase exponentially since the number of new births will be directly dependent on the number of available parents. Human perception of sound levels is designed to change with exponential increase in sound energy. Growth of a savings bank ac-

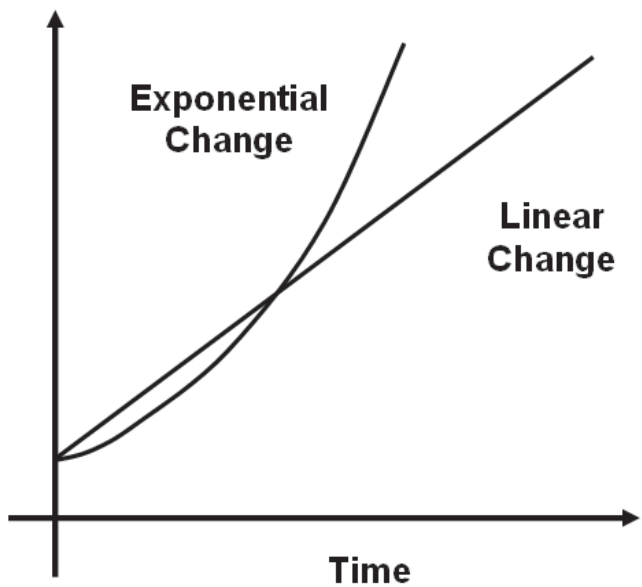

Figure VI.2, Superposition of linear and exponential change

count with compound interest is another everyday example of exponential growth. In figure VI.2, the exponential increase is superimposed by a linear variation. Obviously the exponential variation is different from the linear change though for small enough changes, a linear relation is quite close. Thus, given a small range of data there is no way to logically select the exponential over the linear relation. With more precise measurements, the linear can be rejected when the curve does not pass though the error bars or when systematic deviations are observed. The second possibility is to extend the measurements over a wider range when the discrepancies can help reject the linear relation. This idea of employing a relation in a limited domain is extremely common in physics and we will return to this later.

VI.3 Physics of the exponential dependence

This simple exponential dependence of the intensity of light with the depth of water can be used to clear up one major philosophical misunderstanding. Any high school text book of physics describes light as electromagnetic waves. Such a wave is described as consisting of magnetic and electric fields which are mutually perpendicular and also perpendicular to the direction along which the light travels. In such a description, the intensity of light is given by the square of the electric field strength. It is perfectly possible to describe the experimental results obtained, namely the exponential decrease in intensity with the depth of water assuming that light consists of electromagnetic waves. In this description water is a dielectric medium that resists the changing fields. Then it is possible to show that an exponential variation with water depth is to be expected.

Everyone is aware that there is a mysterious physics called quantum mechanics. According to this quantum description, light consists of particles called photons. The intensity of light is given by the number of photons present. As the photons travel in the water, the electrons in the water molecules absorb these photons and then they are re-emitted. This later can occur in all possible directions and some energy loss is also possible. The consequent lowering of intensity can also be calculated and once again the theoretical description confirms that the variation of intensity with depth of water must be exponential. Thus there is a choice in the physical theory employed.

Consider another physics observation that also confirms an exponential relation. Consider one gram of a sample of C^{14}, a radioactive isotope of carbon with atomic mass 14 amu in contrast to the more abundant C^{12} which is stable. After 5730 years, there will half a gram of C^{14} left. The remaining has been converted to nitrogen. After 11,460 years there is quarter of a gram left. Thus we have an experimental quantity, the amount of C^{14} in a sample varying exponentially with time. However, only a quantum theory can be used for explaining the physics. In the sample, there are about 10^{22} atoms of C^{14} to begin with. About 10^{18} atoms have decayed in the first year. However, one atom may remain un-decayed even after 100,000 years.

This is the essential probabilistic nature of quantum theory. Why should one atom decay while the other does not? Quantum theory can be used to theoretically explain the exponential nature. However, it stipulates that only a probability can be calculated. This led to strenuous objections from Einstein. However, no alternates to quantum explanation have emerged. While all this makes quantum physics very attractive for philosophers we limit ourselves to noting the existence of one example of exponential variation that cannot be explained by classical physics.

As another example of exponential dependence we can consider the measurement of intensity of sound waves instead of the intensity of light. Once again due to loss of energy, the exponential change of sound intensity with distance can be experimentally demonstrated. Physics describes sound as a wave propagation and the observed functional dependence can be understood. The interesting aspect is that there is no simple and obvious quantum explanation for the variation of sound intensity. Only quasi-philosophical arguments that sound waves travel through material medium such as air and that air consists of atoms whose properties are quantum mechanical and that the attraction and repulsion between atoms in the gas are quantum mechanical in nature etc etc can be made. It is true that under certain conditions, in solids at low temperature for example, the quantum behavior of sound (mechanical waves) can be deduced from experiments. However, here is an example of the same exponential relation in physics that has no simple quantum explanation.

The first lesson from the study of functional dependence in physics is that the fundamental theories of physics are employed depending on the experimental situation and there is no automatic preference for more fundamental theories. This is also an everyday experience. Not withstanding the certain knowledge that the earth is rotating round the sun, one refers to sunrise and sunset. In view of the utility of the description the more fundamental heliocentric theory is ignored. This is exactly the situation in physics where, even today, utility leads to choice of physics theories and not their being "fundamental" or "universal". A second lesson is the role of theory for identifying the functional relationships. In the earlier chapter while discussing least square fit, it was observed that the choice of the functional form should be based on arguments other than the data. The fundamental theories of physics provide a solid example of such arguments. In the present example, an exponential would be preferred over the straight line because of the theoretical support. The strength of the theory will be discussed in the next chapter.

VI.4 Domain of functional dependence

Two different types of relationships were discussed in the previous chapter. One involves two uncertain numbers, both of which are considered as equivalent. The second considers one to be an independent variable and the second to be the dependent variable. In an experimental science like physics, the second dominates. The three examples discussed above have either the distance or the time as the independent variable. It is assumed that any possible values of independent variable can be selected. However, even in simple systems that are analyzed in physics, the possible values are limited. We shall discuss a few examples to highlight this limitation.

Let us consider pressure. Pressure for a physicist is the force that is applied on a unit area of surface. The atmospheric pressure is the force exerted downwards by the air. Under normal conditions at sea level, this value is approximately one kilogram for every square centimeter of the surface of the earth. If a fixed quantity of gas in a container is subjected to more pressure, its volume decreases. Boyle's law discovered three hundred and fifty years ago defined a functional dependence with the volume decreasing as the pressure is increased.

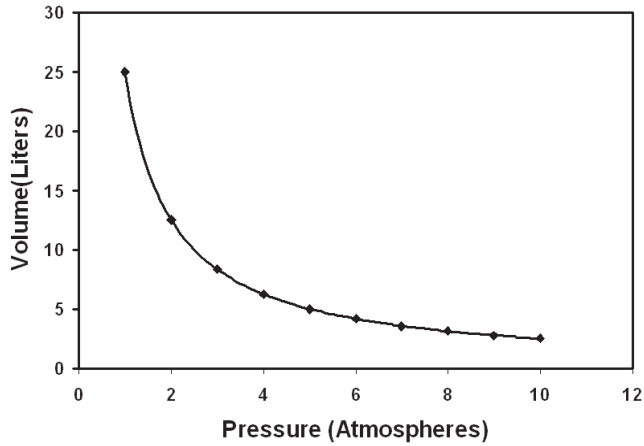

Fig VI.3 Volume of a gas decreases inversely with pressure in accordance with Boyle's law

This is a quantitative inverse relationship and the product of pressure and volume is constant. The simple relationship is shown in figure VI.3. Subsequent work however showed that this constancy is true only if the temperature is held constant. Further, experiments performed at successively lower temperatures show larger and larger deviations from the behavior expected from Boyle's law. First there is a small distortion and eventually the curve breaks into three different segments as shown in figure VI.4.

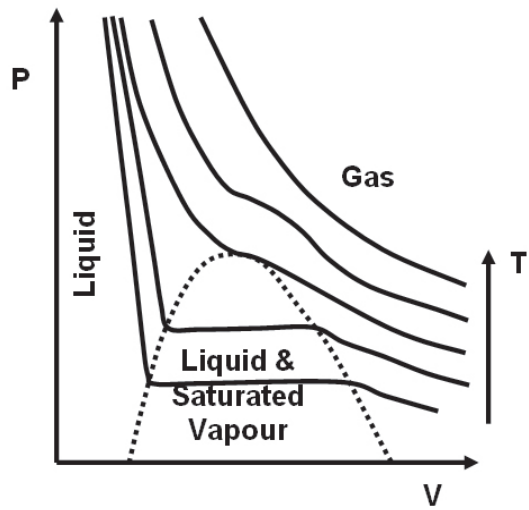

Fig. VI.4 The deviations from Boyle's law for a real gas

At sufficiently low temperature, increasing the pressure on a real gas leads to its liquefaction and not merely reduction in volume. At the singular point in this process, called the critical temperature, extremely interesting physics has been observed. While it is not really relevant for the present discussion, the gas at the critical point shows quite astonishing similarity to the self similarity discussed in the earlier chapter. As everyone knows there is a clear difference between the densities of gas and liquid. At the critical temperature, a drop of liquid contains smaller droplets of the vapor and each of these smaller droplets of vapor in turn has even smaller droplets of the liquid. This is striking evidence of a real system exhibiting self similarity, observed in deterministic mathematical calculations. For the present discussion, the basic lesson from the results is the possibility that an external parameter can influence the relationship and thus a single relationship established under one set of controlled parameters (one gas, one temperature in this example) cannot be the basis for major conclusions.

VI.5 Choice between multiple interpretations and emergent phenomena

History of the physical interpretation of the observations of Boyle is very interesting. Newton had shown almost immediately that if one assumes a repulsive force between stationary particles of gas, volume would inversely depend on the pressure. Pressure would then have to be seen as a repulsive force between the particles and the container. This as we know now is a wrong physical picture. Two hundred years after Newton, Maxwell and Boltzman showed that the particles of gases (by then recognized as molecules) are constantly in motion and pressure is the due to transfer of momentum during collisions from the molecules to the walls of the container. Once again, an inverse relationship between the pressure and volume can be deduced from this picture. An explanation such as of Newton, confined to only one set of observations, is not very useful. The Maxwell Boltzman description is supported by other experimental results making the interpretation more robust.

The example also demonstrates the idea of an emergent property. This is another favorite of the philosophers. An emergent property is attributed to an organism or an entity as a whole and not to its parts.

This is the usual argument against reductionism or reducing the properties of the whole to its parts. Pressure is a wonderful example of an emergent property. It cannot even be defined except as the total effect of a large number of collisions. It is a property of the quantity of gas as a whole. Pressure however is merely the total momentum transferred to the walls of the container by the rapidly moving molecules. Consequently a philosopher would refuse to accept pressure as an emergent property at all and explicitly demands that an emergent property just cannot be summed from its parts.

Interestingly, pressure as a parameter becomes ill defined and non-usable under certain experimental circumstances. One good example is the new technological area of MEMS. This refers to the fabrication of extremely small mechanical devices such as gears, cantilevers and wheels using the technology developed for electronic chips. Devices as small as a few parts per million of a meter are routinely produced. At these dimensions, in some designs, the number of collisions becomes small, the momentum transferred by individual collisions changes the device and there is no average value of the pressure. But this practical reality does not convince those committed to holism and the idea of irreducible emergent properties.

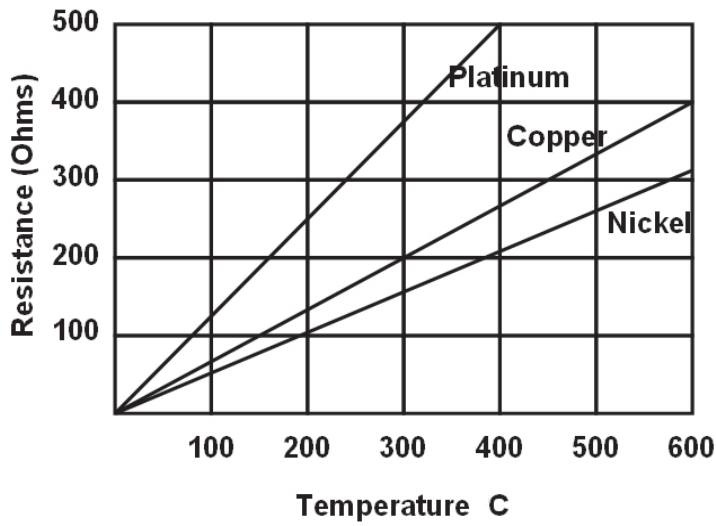

Figure VI.5 Variation of the resistance of copper, platinum and nickel with temperature

VI.6 Surprises during extrapolation and interpolation

As another example of experimental functional relationships observed in physics, we consider the variation of the resistance of metals. The typical variation for three metals, platinum, copper and nickel is shown in figure VI.5. While the curves look linear, the variation is actually quadratic. Resistance $R = a + bT + cT^2$ is the equation employed. Here T is the temperature and a,b,c are constants. The relationship holds for many different metals. A simple quantum mechanical theory that considers electrons in these metals as free particles can predict the observed variation of resistance.

Temperature is the independent variable. Changes in temperature as small as a thousandth of a degree are routinely measured by measuring the corresponding change in resistance of a platinum resistance thermometer. Thus it appears that the basic description offered by the physics is acceptable. However, consider the resistance of mercury shown in figure VI.6. Over a wide temperature range, the resistance of mercury as with other metals can be described by a simple continuous curve. As temperature is lowered below 4.15K, (-269C), the resistance abruptly goes to zero as mercury becomes a superconductor. If electrons in metals are free particles then the resistance cannot become zero. That theory cannot predict the sudden change that is observed on extrapolating the measurements made above the transition. A more sophisticated or complicated theory

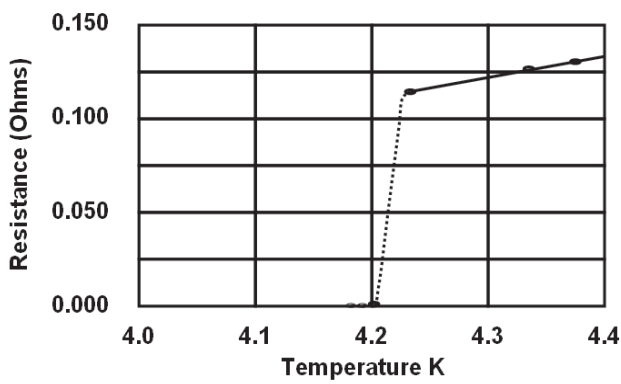

Figure VI.6 Variation of the resistance of mercury showing the superconducting transition at 4.15 K (-269C)

that includes the interactions between the electrons and the rest of the mercury atoms becomes necessary. The possibility of this failure or limitation of the free particle model is not predictable from the high temperature measurements. The approximate model continues to be employed despite the failure to predict all properties. The key issue in all cases is the necessity for a more accurate description. An example of failure during extrapolation was presented in the last chapter. The present example shows that such problems emerge even in basic physics.

Consider the resistance of nickel measured more accurately in a small temperature range of 350-380C that is shown in figure VI.7. If measurements are done at large temperature steps, for example if data is obtained only at every 20C, the small deviation between 350-380C could be mistaken as an error in the data points. Interpolation would fail to predict the small change at 358C. However this small change is the result of nickel changing from a ferromagnetic to paramagnetic form at a specific temperature. When the curve is differentiated to get the temperature coefficient of resistance, there are dramatic changes at the critical temperature as shown in figure VI.7. While the description of electrons in nickel as free particles is justified both above and below this temperature, the approximation breaks down near the transition. The behavior in that region is very interesting. A paramagnet can be imagined for simplicity as a

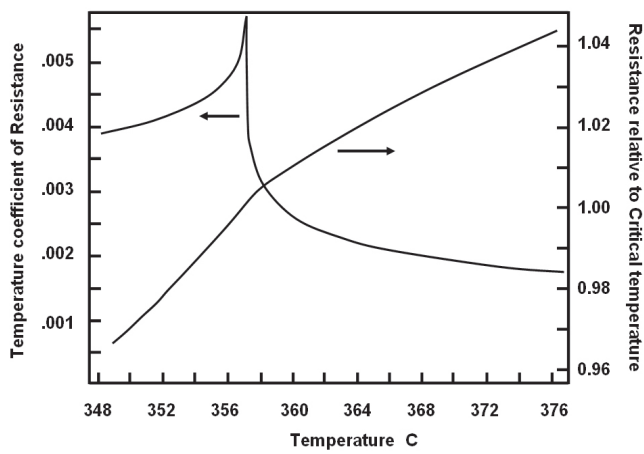

Figure VI.7 Variation of resistance and temperature coefficient of resistance (the differential) with temperature

collection of small magnets each of which point to a random direction, unlike the familiar magnetic compass which invariably points only to the north. Similarly the ferromagnet is a collection of small magnets but they all point in the same direction. A compass is made of a ferromagnetic material, which is free to move and since the earth itself is a magnet, points north. Most interestingly, the same self similarity we described in the gas liquid transition is observed here. At the critical temperature, small paramagnetic regions exist inside the ferromagnetic region and smaller ferromagnetic regions exist inside these paramagnetic regions. Now all this beauty is lost and not appreciated under an approximate theory. That incidentally is the correct way to look at extrapolation and interpolation in physics. An approximate theory is not wrong. It simply fails to make a more accurate and often more beautiful prediction.

VI.7 Failure of idealization and definition of limits

Even this cursory examination of physics as an experimental science reveals many interesting consequences of the effort to quantify measurements and use it to understand "how well we know it". Establishing mathematical relationships between experimental data is relatively easy. With the availability of computers, fitting experimental data to many mathematical forms has become almost trivial. Caution is advised in drawing conclusions from these mathematical exercises. The observed functional dependence could be completely lost by changes in an as yet unknown parameter. That Boyle's law was not valid near the critical temperatures took more than two hundred years to realize. Not withstanding this, there are large areas of science and engineering ranging from climate modeling to pneumatic engineering where the simple law of "ideal" gases is still employed. Similarly, both extrapolation and interpolation are risky. The key issue is to recognize that the association between a mathematical variable and the underlying physical reality is always only approximate. When the limits of the approximation are well defined, practical utility within the limited regime is not compromised.

———————

VII
HOW RELATIONSHIPS ARE UNIFIED IN PHYSICS

VII.1 Relating multiple experimentally observed relationships

The brief discussion of physics in the previous chapter introduced the idea of theories of physics. Both classical and quantum theories were briefly mentioned. The key for the development of such theories is the relation of one set of experimental observations with another. In the earlier example, Boyle's law is valid when the temperature is both higher than the critical temperature and is constant. When the pressure, volume and temperature are allowed to vary a new law called the ideal gas law is obtained. In social sciences too there are often a number of variables that influence the one which is being observed. For example, the salary earned by an individual could depend on education, experience, Intelligence Quotient (IQ) and (regretfully) gender or race. The teasing out of the relative importance of each of these contributors is accomplished using an extension of the correlation procedure described in the last chapter called multivariate regression. However, there is a fundamental difference between multivariate regression where the relationships are all defined by correlation coefficients and theories of physics as discussed below.

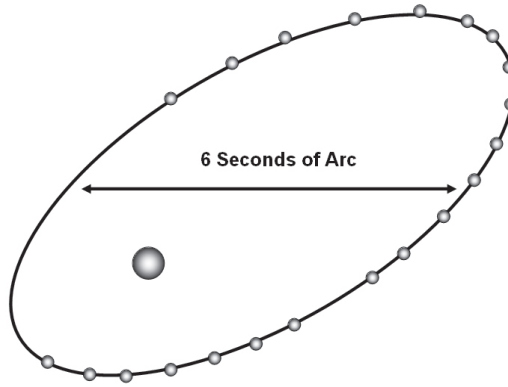

Fig. VII.1 Experimental observation of the orbit of a binary star

Consider the observation of a binary star system shown in figure VII.1. Unlike the example chosen in the first chapter this one is a perfect ellipse with the massive star at the focus. Once again the actual data has been traced ignoring some details not relevant here. The observed data points can be represented by the mathematical formula for the ellipse, $y^2/b^2 = 1\text{-}x^2/a^2$. All planets move in an elliptical orbit with the sun as one focus is the first law of Kepler in physics.

Now consider another set of experimental data showing the variation of the acceleration due to gravity with altitude. This quan-

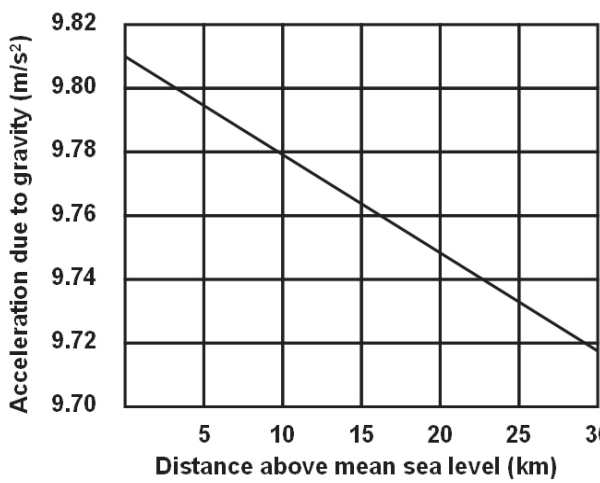

Fig. VII.2 Variation of the experimentally determined acceleration due to gravity with altitude

tity is commonly determined by physics students in the high school using a stop watch and a simple pendulum. Figure VII.2 shows the value of the acceleration due to gravity which is also equal to the gravitational field strength measured as a function of the altitude above mean sea level. The acceleration due to gravity decreases as the square of the distance from the center of the earth, which is altitude above sea level added to the average diameter of the earth. However, the altitudes one considers are a very small fraction of the earth's diameter. For example, data up to 30 km are shown in the figure, much smaller than the radius of the earth, 6371 km on an average. Thus the curve is experimentally very close to a straight line.

A little bit of mathematics will enable one to realize that the two mathematical relationships, the first that defines the orbit of a planet as an ellipse and the second relating the acceleration due to gravity to the altitude are equivalent. In other words, if one of these is true the other is also true. If there is a force which (in physics force is simply the product of the mass and acceleration) decreases as the square of the distance then the acceleration will decrease as seen in figure VII.2. Also the movement of a satellite will be along an ellipse. If elliptical motion has been observed, it can be deduced that the force decreases as the square of the distance. However, this inevitability is only apparent in the light of the mathematics.

The movement of planets in elliptical orbits provides another example. Such mathematical linkages in theoretical physics are considered "fundamental" by physicists. The second of Kepler's laws states that "a line joining a planet and the sun sweeps out equal areas during equal intervals of time". In figure VII.3, the shaded regions are equal according to this law which has been verified experimentally. The distance between the sun and the planet is not constant since the orbit is not a circle with the sun at the center. In order to sweep equal areas in equal times, the velocity of the planet changes, being higher when the planet is nearer to the sun. Thus we have another relationship that is implied. The velocity of the planet and its distance from the sun are inversely related. This mathematical relation can also be derived with a little bit of mathematical skill if it is assumed that the force varies inversely as the square of the distance. The product of the mass of the planet, its distance from the sun and

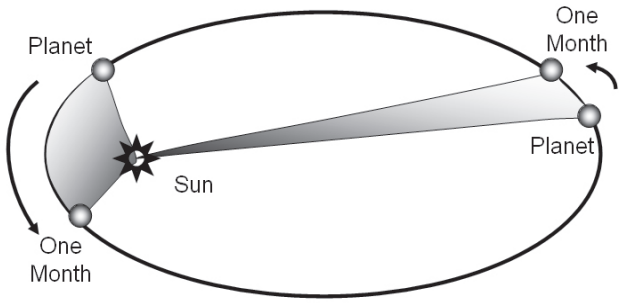

Figure VII.3 Kepler's second law states that the shaded regions are of equal area

the velocity is called angular momentum. Thus Kepler's law is equivalent to a statement that the angular momentum is conserved or held constant.

Experimentally the constancy of the angular momentum in the case of the planets is approximate. The earth's rotation is observed to be slowing by about 1 second every 50,000 years from the current value of 24 hours. This would straight away imply a reduction of the angular momentum. The moon is observed to be moving away from the earth at an average rate of about 3-4 cm per year. The total angular momentum including contributions from the rotation of the earth and the moon is conserved more precisely than of the earth alone. The reduction of the earth's angular momentum due to the slowing down of the rotation is compensated by the increased earth moon distance. While experimentally investigating the conservation of the angular momentum of the earth, in addition to the forces due to other planets and the moon, friction due to tides has to be taken into account depending on the accuracy desired.

As in the example briefly cited in the first chapter, about the binary star, every effort is made to understand the reasons for the observed lack of constancy of the angular momentum rather than accept its failure. The reason for this becomes clear during the course of the discussion in the next few sections on unification of relationships in physics.

There are a huge number of experimental observations which show that in the absence of external forces, angular momentum is

always conserved. Technically torques (a force applied at a distance) should be absent. For example, the force applied at the edge of a door becomes a torque at the hinge. Not only is angular momentum conserved for a planet moving around the sun it is also conserved for an electron moving inside an atom. Thus one gets relationships that are called more fundamental since they are obeyed by more and more diverse phenomena.

Such experiences have lead to statements like "The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve" (Eugene Wigner) or "How can it be that mathematics, being after all a product of human thought which is independent of experience, is so admirably appropriate to the objects of reality?" (Albert Einstein). Clearly, there is no possibility of this great wonder being resolved in the present discussion. But the goal of the present approach is very limited. It is to first demonstrate the fundamental approach of this mathematical formalism and then to explore how these overarching fundamental theories have to be employed in understanding physics.

VII.2 The three equivalent models of force, local field and least action

The above description gives a flavor of situations in physics where several experimentally determined relationships can themselves be related through mathematics but not otherwise. Before we discuss the strength of such mathematical unification, which is what one means by a fundamental theory of physics one has to appreciate the redundancy in such descriptions. We discuss three possible theoretical descriptions of the same experimental observation. All of them are mathematical and are also interrelated through mathematics. The key question is whether any one of them can be identified as more "fundamental".

Newton's law of gravitation proposes that a force of attraction exits between any two material bodies and that the force is proportional the product of the two masses and inversely proportional to the square of the distance between them. This is called the action at a distance

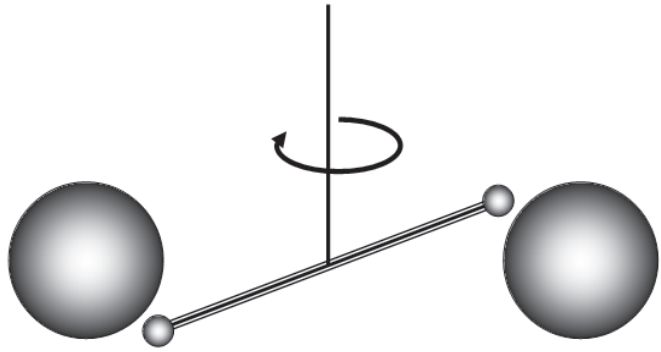

Figure VII.4 Experiment to determine Gravitational Constant and thereby weighing the earth

description of force in classical physics. Consider the simple experimental setup in figure VII.4, in which, two balls are suspended from a thin fiber and then brought close to two other balls. The force of attraction causes the fiber to twist as shown in figure. This twist can be measured and the actual force determined and the constant of proportionality, called the Gravitational Constant calculated. Cavendish who performed this experiment for the first time called this "weighing the earth" since from this experiment the weight of the earth can also be calculated. This then is one mathematical model (called action-at a distance picture) that can be used.

The easy question to ask is why there is a force such as gravity? This uncovers a standard science problem. When a phenomenon is explained newer more interesting questions emerge to challenge succeeding generations. Einstein answered this challenge by showing that gravity is the consequence of local deformation of space-time, thus permitting another question to be asked, why does that happen? Ignorance always remains in science; it is only reduced.

A more difficult question is "how does one ball know the position and direction of the second?" If the mass of the second ball were doubled, the force would also be doubled. If the distance were doubled, the force would decrease by a factor of four. But the force is exactly correct as if the first ball "knew". This "knowledge" question troubled the scientists of the age. It is somewhat weird to expect that an inanimate object "knows".

A new mathematical formulation, developed a hundred years or so after Newton considers not the force, but the path along which bodies move under the influence of the force. For example, the earth moves around the sun in an ellipse. It is possible to mathematically show that a body acting under the influence of an attractive force such as described above would move in an ellipse. But it is also possible to determine the potential and kinetic energies of a body and calculate their difference which is called "action" in physics. The path followed by a body between two points in space, say the position of the earth on two different dates is that along which the total action is minimum. This is called the path of least action in the path formalism of classical physics and is mathematically equivalent to the force description. Now the body in motion does not need to "know" the position and direction of the second object but perhaps has to "sniff around the path it follows", determine the path of least action and then follow it. This is also weird but now the weirdness has changed!

A third formulation follows from the existence of a field. One assumes the existence of a field of influence around every body. The second body responds to the field present at its location due to the first. The field present can be calculated from local measurements without knowing anything about the first body. One imagines a sphere around the second body and the local potential, defined as the average potential on the sphere can be determined. Then knowing the mass present inside the local field can be determined mathematically. In this local field picture of classical mechanics, the potential, field and hence the force are local properties. This description appears to be a better view of the physics and more fundamental. At least Einstein thought so. The advantage was not merely the different mathematics.

The concept of field did not have any "measurable consequences different from action-at-a-distance view of gravity". But it eliminated the weirdness of a how a body could know the location and mass of other bodies. It does not need the body to sniff around the path it travels. The ignorance remains. Why a force of gravitation exists was not answered. Why a field exists is also not answered. It should be clearly recognized that the three descriptions, force acting at a distance, path of minimum action and local field defined from a potential are all mathematically equivalent. The choice

by a physicist depends purely on the ease of solving a particular problem.

The definition of the field enabled more fundamental questions to be asked in physics. This enhanced the depth of understanding. The field would act at all points of space, in principle even at infinite distance. Can the change in field be communicated to an infinite distance instantaneously? Solving this question resulted in Einstein's theory of relativity. "Gravitational waves" convey changes in the field strength with a finite velocity.

Actually most lay people or for that matter physicists do not just stumble on to the problem with the Newtonian action-at-a-distance. I am not aware that any student has actually taxed a middle school teacher with such a question. Scientists simply did not decide to "forget" about it. Scientists would only think about a problem if they can come up with a meaningful resolution. We do not know how many people might have been "concerned" about the problem but only Einstein could provide a resolution.

There is nothing fundamental to choose between the three different ways of formulating the theory. An interesting way of explaining this absolute equivalence was described by Feynman. In the standard formulation of geometry, a set of axioms are takes as the starting point. The specific axioms of plane geometry were described earlier. Also the Pythagoras theorem was described. It is in principle possible to start with the theorem as an axiom and derive the usual axioms about the nature of a straight line and point from it. After all only the logic is being reversed. Thus there is nothing "sacred" about the axioms. They are selected for their utility in explaining various geometrical theorems. Similarly the choice between the various descriptions is based on utility.

The weirdness of the body "knowing" the existence of another or a moving body sniffing around the path were deliberately mentioned. Concepts of physics when described in ordinary language can create "weirdness". This weirdness has become quite notorious in the case of quantum theories and has resulted in great philosophical debates. We shall return to this a bit later in this chapter.

VII.3 Why are nature's laws the way they are

The above description shows how nature's laws are written in mathematics as was first clearly stated by Galileo. The experimentally determined quantities such as acceleration can be treated as abstract mathematical entities and related to other experimental quantities such as the velocities of planets in their orbits. There are multiple descriptions possible, all of which are related through mathematics.

More relevant to the present discussion is the nature of acceptable mathematical structures. As Vic Stenger says "what are the restrictions on the way physicists can formulate their mathematical statements about observations?". The key requirement obviously is that the predictions made on the basis of the mathematics should be specific enough that they can be compared with observations. The whole of science rests on the ability to predict an outcome on the basis of a theory and then conclude that if the experiment does not confirm the prediction, the theory is to be rejected or modified.

An experiment that cannot be repeated is useless. Thus it is necessary that the experiment be repeatable anywhere. Shifting the experiment to a new location or changing the orientation should not change the outcome. Similarly the experiment must be repeatable anytime in the future. However, even a simple pendulum will not have the same time period when it is moved about on the earth. A local change in the acceleration due to gravity can change the results. By observing the changes systematically, it is possible to realize that in such cases the experimental setup is influenced from outside. Here the non-local influence is the earth's gravity. Influences not confined to the experimental setup are transmitted to the experimental setup through physical processes.

As a thought experiment, if the earth's gravity suddenly changes by a large amount, the changes should be physically transmitted to the pendulum. Obviously if such a change is physical, it must have a finite velocity. To summarize , an experimental result should have (i) A small number of possible preferably quantitative predictions, (ii) Constancy of the experimental results when repeated

at a different location in space and time and (iii) A finite velocity with which non-local changes influence the experiment.

Thus the theory that is developed must include these features. The predicted value of the experimentally observed data point does not depend on where the experimental setup is located or the time of observation. Further, all influences external to the experimental travel with finite velocity. It can be mathematically shown that conservation of momentum is the consequence of the demand that the experiment performed at a different location show the same result. Similarly, the law of conservation of energy is the result of the demand that the experiment be repeatable after any delay in time. The rotation system by a constant angle results in the law of conservation of angular momentum. That changes in non-local influences should take a finite time to reach the experimental setup is essentially the special theory of relativity, which showed that the maximum velocity for the transfer of information is the velocity of light. In the language of mathematical physics these are called continuous symmetries and thus any mathematics must accommodate these symmetries. In the absence of these symmetries, equations of physics do not describe objective reality. The principles of conservation of energy, linear momentum and angular momentum, and special relativity are required for physics to describe objective reality. Once there is a force, momentum is not conserved since force is rate of change of momentum. Or if momentum is observed to change, the change can be predicted mathematically in the language of a force. Obviously the various equivalent mathematical theories that are proposed to explain the results of experiments do have a restricted range of possibilities. The laws of nature which we have are in the form they do have purely because they are expected to make predictions that can be compared with experiments.

VII.4 Limits on mathematical structures

In addition to symmetries as outline above, there are other limitations on possible mathematical theories. Symmetries are primarily requirements of fundamental physics. Though Feynman joked that “the name fundamental physics was stolen by us to give the other physicists an inferiority complex”, it is nevertheless true that only in this area are the symmetries explicitly taken into

consideration. The appreciation of these symmetries makes some really great physicists wax lyrical about the beauty of these equations. However, beauty is not the key issue. A beautiful equation that does not predict observed experimental phenomenon is useless. However, absence of the critically required symmetry enables outlandish ideas to be rejected out of hand. Unfortunately proponents of "crackpot science ideas" never understand that the scientists have a short cut like this for rejecting them.

In an earlier chapter we talked about classical and quantum explanations. We now try to superficially understand the requirement for both these classes of theoretical explanations. There is another important distinction that is not familiar to the general public despite the great fame of Einstein. These are technically called relativistic and non relativistic theories. The interrelationship between all these is very instructive to understand how well we know things in general.

Among the symmetries mentioned above, the symmetry introduced by the special theory of relativity, called the Poincare symmetry is singularly important. Classical and quantum theories have been developed with or without Poincare symmetry. We need to understand why? The physics describing the movement of planets in their orbits is a classical non-relativistic theory. However, the movement of the perihelion of planet Mercury (this was mentioned right in the first chapter) requires a relativistic classical theory. This is the Einstein's theory of general relativity. The most important limitation of such "better", "correct" or "new" theory is that the older data that were being explained on the basis of other theories have also to be explained. The general theory of relativity does not merely explain the movement of Mercury. It also explains the motion of other planets. Since the mathematics is more complex, the non-relativistic theory continues to be used in practice. This is similar to the use of the terms "sun rises" and "sun sets". Actually the earth is moving but for the purpose on hand, a description assuming the sun moves is more convenient.

Just as gravity is the force of attraction between large objects, the electrical forces of attraction and repulsion are observed. Classical relativistic theory of electric fields is used to describe many

experiments. The requirement for quantum theories emerges when the electrical forces at very small distances, such as within an atom are to be understood. Here, the deviation between the predictions of classical theory and observations are large and there is no classical description. In a restricted set of experiments, usually at large dimensions and large amounts of charge, the classical description is satisfactory. A more "difficult" or cumbersome quantum description can be provided. As before, the relatively more recent quantum theory is coherently integrated with the earlier theory of classical electromagnetism. This is not because the theories are "sacred". It is because the existing theories had predicted the outcome of some experiments and the new theory has also to do the same.

The question of weirdness, either the ignored classical variety mentioned in an earlier section or the more famous (infamous?) counter intuitive weirdness of quantum theory can be well appreciated with an analogy from geometry. Geometry, as has been said, begins with a small set of axioms (the definition of a point, a line, a straight line and parallel lines etc). All the theorems of geometry for example, the Pythagoras theorem follow from these axioms through logic. Similarly we have the assumption of particles and fields and the consequences of the experiments are deduced based on these assumptions. The assumptions regarding the fields in classical theory are like the simple (Euclidian) geometry. The assumptions (axioms) seem to be sensible. On the other hand the assumptions made regarding the quantum particles seem to be "weird". To be sure there are many things in advanced mathematics which are also weird. But a mathematical concept no matter how weird is "imaginary" and so this causes no problems for the general public and no philosophical interest. The "weird" quantum object is "real". This leads to huge impact on general public as well as philosophers. It also leads to "crackpot" science that has only the justification of being weird. This problem is not easy to solve. It is difficult to simply reconcile to the reality that if we want the results of our mathematics to agree with experiments, the concepts "have to be weird".

One other fact has also to be admitted. While the quantum theory has been able to make predictions to astounding accuracies, in addition to the weird assumptions there are doubts even among

physicists whether all steps in the mathematics are absolutely justified like the steps leading from the axioms of geometry to Pythagoras theorem. However our aim is more modest. We simply ask how well we know. The question is empirical so predictability is satisfactory and philosophical conundrums are unimportant.

VII.5 Two extreme examples of mathematical theories

The restrictions on possible mathematical theories in the light of the above description are exemplified by two examples. The first is theory of superstrings. Even a cursory look at the theory is irrelevant for our purpose. It has been actively researched for nearly three decades. As things stand today there is no quantum theory of gravity and the only working theory, Einstein's general theory of relativity remains a classical theory. The rest of fundamental physics is described by a quantum theory and classical electromagnetism is useful only in a limited area. Superstring theory is supposed to bridge classical and quantum theories leading to a "theory of everything". There has been an extensive mathematical effort that is still ongoing but the key worry for most physicists is the lack of experimental support. There are no current experiments that can "only" be explained using string theory and more worryingly most versions seem to have no testable experiments that are even within the limits of human capability. What to make of such a theory has been vexing physicists for some time. This is an illustration that detailed, elaborate and complex mathematical theory can still be of questionable use. This difficulty must serve as a warning to ridiculous amateur attempts at grand unified theories which are obviously a case of fools venturing where angels fear to tread.

A second example is from the domain of the amateur arguments. Radio isotope dating shows that the ages of certain rocks on the earth are as high as several billion years. This is not quite palatable to some for religious reasons. One of the interesting approaches in these circles is to suggest that maybe the half life of the elements is not constant and thus we get these large ages. This example is relevant in the current context in illustrating the consequences of ignoring mathematical linkages. The exponential variation of the number of atoms undergoing radio active decay is not an isolated

fact. This is mathematically related to issues such as tunneling of particles, strength of the weak interaction, and so on. Thus permitting the lifetime to be a variable rather than a constant may help in explaining away the large ages attributed to rocks but this would make consequential changes in predictions regarding many other phenomena. These predictions would contradict observations and thereby make it clear that such ad hoc adjustments in one part of the theory are not permissible in physics. This is a direct example of the strength of fundamental theories. It is the existence of mathematical linkages that prohibits ad hoc tinkering with individual aspects of the theory. If the relationships were purely empirical there would be no way of questioning such attempts.

VII.6 Approximations and phenomenological models

The above discussion on fundamental theories of physics should be tempered with an understanding of the actual situation. The first limitation that has to be acknowledged is that all experimental knowledge is only approximate. Thus a planet such as the earth travels around the sun only in an approximately elliptical orbit. The forces caused primarily by the moon and to a lesser degree by the other planets change the ideal elliptical orbit. Further the water in the seas is also influenced by gravity resulting in tides and the tidal movement results in friction and loss of energy. Thus as more and more precise observations are made, more deviations can be identified and then accounted for by more precise mathematical modeling. This continuous improvement gives confidence in the system. It has also to be noted that some of the corrections turn out to be phenomenological. This term which shall be used extensively in the later chapters indicates that a precise theoretical calculation is not possible. In the above example, while the contributions from the moon and the planets can be calculated theoretically using the laws of gravity, tidal friction can only be estimated without a basic theoretical framework. As another example consider the absorption of light in liquids discussed in the last chapter. It is possible to obtain a quantum or classical theoretical explanation for the observed exponential decrease of intensity. However, the actual magnitude cannot be defined from theory. Consider that one adds a colored salt like copper sulphate to water. The variation with depth of water will be exponential for all

concentrations as long as all the salt is dissolved. However, the variation of the amount of light with the amount of copper sulphate will follow a mathematical curve that can be only empirically observed. Thus, in the context of physics, there are fundamental theories, approximations and phenomenological descriptions all coexisting at every level. While it is possible to make theoretical predictions regarding the evolution of the universe a few seconds after the big bang that has created the whole of the universe, there is no theoretical ability to calculate from the properties of water and iron, the amount of water that will flow out of an iron tube of given length when connected to a tank of given height. The planning for plumbing connections in the house are made based on empirical rules.

VII.7 Devil is in the detail

Trying to tease out order in physics and separate it into fundamental and empirical is much more difficult than to identify the regions of chaos described in an earlier chapter. The devil is in the detail. It is important to understand the strengths and weaknesses of physics. That alone will enable a proper appreciation of how to evaluate the strengths and weaknesses of knowledge in other areas and that in turn will enable us to answer questions about how well we know things. In the next chapter a really solid example of physics and engineering is discussed in some detail. That would enable a proper appreciation of the problems and the correct methodology for analysis. Railing at "physics approaches", philosophizing based on physics concepts and accusations of physics envy are all equally useless.

VIII
HOW FUNDAMENTAL THEORIES CONSTRAIN THE REST

VIII.1 An example from physics and engineering

Examples involving fundamental theories such as quantum mechanics and Newtonian gravitation in addition to more empirical observations such as Boyle's law were discussed earlier. The key aspect of knowing how well we know anything is to quantify the confidence. In reality this confidence is not uniform across physics, engineering or for that matter even a small practical issue. Rather than philosophical arguments about the limits of what science can and cannot do, the present part seeks to see what lessons physics has to offer and to that end, a specific example is being chosen as an illustration. The example is a device called a metal oxide semiconductor field effect transistor or a MOSFET. There are millions of these devices in every computer. The Integrated Circuit industry is relentlessly trying to make smaller and faster transistors. This improvement is at the heart of the faster and faster computers that are being produced. The operation of the device is rather simple. A MOSFET has three electrical connections and acts typically like an electrically operated switch. A small voltage is applied to one connection called a gate. This permits a current to flow through the

other two called the source and the drain. When the voltage is removed there is no current flow. There are other versions of a MOSFET which operate in reverse, that is to stop a current when a voltage is applied to the gate. Also there are situations where the gate voltage continuously controls the magnitude of the current rather than switch it on or off. But we will consider only the first variety.

VIII.2 Concept of the MOSFET

A MOSFET is a glorified switch that can be fabricated at dimensions smaller than a millionth of a meter with a current flowing through the gate being less than a billionth of an ampere, the ratio between the current from the source to the drain to the current in the gate being a million and the time taken for switching being a billionth of a second. Such numbers are possible only because of the peculiar concept of the MOSFET.

As the name implies the device consists of a metal, a semiconductor and an oxide which is an insulator. The common examples could be silicon as a semiconductor, aluminum as the metal and silicon dioxide as the insulating gate material. The schematic is shown in figure VIII.1. Why silicon is a semiconductor and how such a material behaves under various conditions are simply impossible to understand unless one uses quantum mechanics. Even the properties of a metal cannot be clearly understood unless one thinks of the quan-

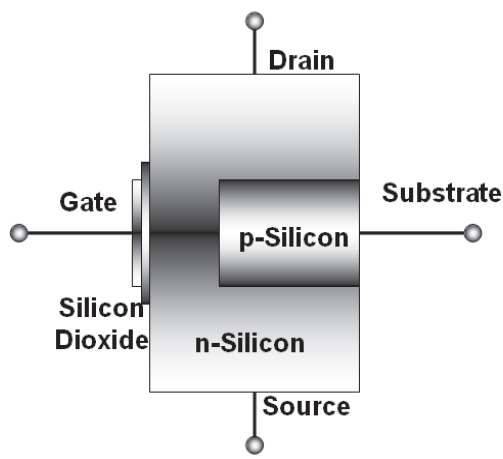

Figure VIII.1 Schematic of a MOSFET

tum mechanical properties of electrons. Even after the electron was experimentally discovered by J J Thomson and Rutherford showed that an atom had some kind of a planetary structure with the electrons moving around the nucleus, the great physicist Lord Kelvin argued that since the free electrons which conduct electricity through the metal came from the individual atoms, they should all be recaptured at low temperatures and the metal should stop conducting. As mentioned before, this is experimentally wrong. The resistance of metals decreases at low temperatures and metals like mercury even show zero resistance rather than zero conductance. The reason of course is that quantum mechanical properties of the electron were not understood till much later.

Further, only quantum description shows that silicon has a small "band gap" while the oxide has a much higher "band gap" and does not permit the electrons to travel into the gate at will. It is a quantum mechanical description that enables one to understand that semiconducting silicon can be both "n" and "p" type and that its resistance changes with the "doping concentration" (the amount of phosphorus or boron that is added to the silicon). The bands "bend" at the interface between the oxide and the semiconductor. "Traps" or defects at the interface control the amount of time taken for the switch to go from a closed to an open position or vice versa. Details of the quantum mechanical descriptions are not important for the present. Classical theories that describe the motion of an electron as a small particle being attracted or repelled by electrical fields cannot simply lead to the concept of a MOSFET. So we simply accept that conceptualizing a MOSFET involves learning quantum mechanics and presumably so does making the device in practice.

VIII.3 Designing a MOSFET

After the concept, the MOSFET has to be designed. For this the various dimensions and properties of the metal, the oxide and the semiconductor have to be specified. A more detailed schematic view of MOSFET is shown in figure VIII.2. This provides views of the device from the side and the top. There are many detailed parameters that are considered by the engineer while he designs the device. Some of the simple ones are (i) the resistances of the silicon in the source,

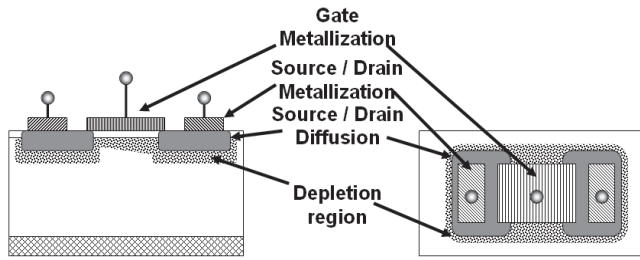

Figure VIII.2 A more detailed schematic of a fabricated MOSFET device

drain and substrate regions which in turn specifies the dopant atom concentration in those regions, (ii)the dimensions of each of the three regions and (iii)their positions with respect to one another. Today there are very detailed models of MOSFETs and simulation packages that help the design. When this design is properly understood and implemented, the engineer should have a reasonable idea of the performance of the device. For example, the speed with which the device responds, the power required for operation and the increase in temperature during operation would be known.

For the purpose of our discussion there are once again only two basic issues that need to be remembered. The electron as described in these models does not have any of the strange quantum mechanical properties. In these models, the electron is a simple charged particle that moves under the influence of an electric field. The band diagrams are converted into equivalent electric fields. Some necessary quantum mechanical complexity of the electron is taken care of by simply assuming that the mass of the electron, designated as an effective mass is different from the true universal experimental value. A vacancy of an electron is treated as a positively charged particle called a hole with its own effective mass. The description is simply one of classical electromagnetic theory mixed with some purely phenomenological arguments. The design of an improved MOSFET never explicitly considers the quantum mechanical theories.

This idea of treating an electron as a “classical” particle in describing model electronic devices has become so widespread that it is not uncommon to see trained device engineers make silly mistakes

when proposing completely novel devices. The recent explosion of research in nano-materials and nano-devices comes with the strange claim that "quantum mechanical" effects begin to be important at dimensions less than about 50 nm (one nm is a billionth of a meter). What is actually being said is that the modeling based on a quasi-classical approximation fails. All electronic devices by their very nature are based on concepts of quantum mechanics.

VIII.4 Fabricating the MOSFET

A cursory glance at the fabrication of a typical silicon MOSFET device would also be instructive. The process where by these devices are fabricated is extremely automated, complex and expensive. Only three steps, simplified to a large degree have been identified for discussion here. The first of these is diffusion of impurities in the semiconductor (Si) to alter the resistance of the various semiconducting layers in the device. Typically the quantities required are one atom of the dopant such as boron or phosphorus for about a million atoms of silicon in the active region, the so called channel region and about one atom of the dopant for every hundred thousand atoms of silicon in the contact regions. Neglecting the technological challenges involved, the process of doping is simple. A relatively large amount of the dopant is placed on the surface of silicon and heated to a high temperature for a long duration. The dopant diffuses into silicon. The actual concentration of the dopant atoms

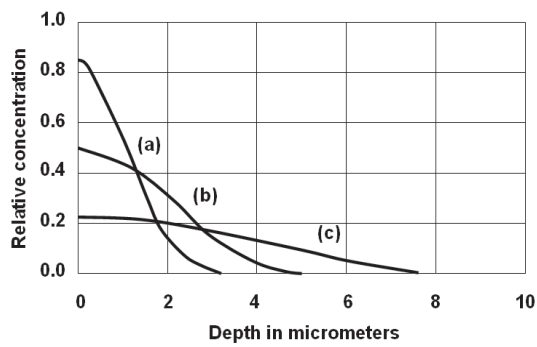

Figure VIII.3 Concentration profiles of dopant in silicon when diffused for increasing time intervals (a) to (c)

inside silicon will resemble the curves shown in figure VIII.3. The mathematical relationship between the concentration and depth involves only one constant, the diffusion coefficient which depends on the semiconductor and the dopant being used. The diffusion coefficient is experimentally determined. This is used to calculate what temperature and time are required to get the required concentration of the dopant and consequently the required resistance.

The physics of diffusion of dopant atoms in silicon is assumed to be similar to diffusion of ink in water or smoke in the atmosphere. Such a simple calculation is sufficient though it ignores even the basic atomic structure of silicon let alone more complex theoretical ideas. The atoms in silicon are at fixed positions and at fixed distances to one another in each direction, forming a crystalline lattice and not moving around freely as in a liquid or gas. For sake of completeness, it should be mentioned that in the case of the latest extremely small devices, some additional complexities such as the size of the dopant atom, whether the diffusion occurs along vacancies or interstitial sites etc are taken into consideration. But even here, these are qualitatively described and then the pure empirical description is suitably used for calculations.

The thickness of the gate oxide is very critical for the proper operation of the device. Oxygen reacts with the surface of silicon to form the first layer of oxide though which oxygen has to diffuse to react with silicon at the interface and increase the thickness. The

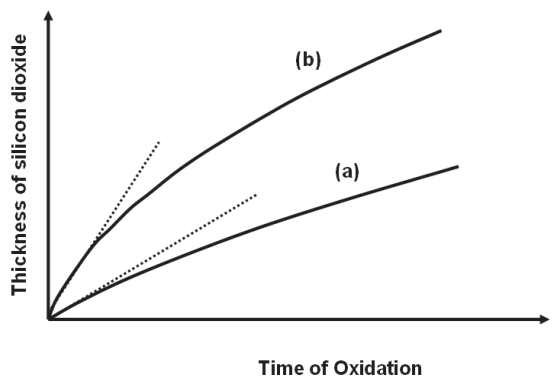

Figure VIII.4 Thickness of oxide grown on silicon at two temperatures

thickness of oxide grown when silicon is kept at a constant temperature initially varies linearly with time and subsequently varies as the square root of time as shown in figure VIII.4. The thickness of the oxide grown will also increase with temperature. In the figure, curve (b) represents growth at a higher temperature as compared to curve(a). Using these experimental results, the time, temperature and source of oxygen are selected to grow the desired thickness of the oxide. For very thin layers there are complications but this simple picture ignoring the details of the chemical nature of oxygen, the atomic structure of the growing oxide and other such theoretical ideas is sufficient. Once again the treatment required is purely phenomenological and is correlated with the quantum physics very superficially. While in the case of the dopant diffusion, the mathematical description is common for a number of dopants and semiconductors, in the case of oxide growth, the models described above are useful only for the case of silicon and oxygen.

As a third example consider lithography. The lateral picture of the MOSFET in figure VIII.2 shows distinct small regions that form the gate, source and drain. The smallest of these dimensions is usually referred to when computer chips are designated as "x.xx" micron technology. Lithography is the process of creating these small regions. We ignore once again a lot of technical details in describing the complex process. Light falls on silicon though a mask onto the required regions. Usually the silicon is coated with a photo resist, a material that becomes soft under the exposure to light and thus gets selectively removed by etching. The physics which defines how light passes though masks was developed by Lord Raleigh and is purely classical physics which assumes that light is wave. The original results by Raleigh show that the wavelength of the light used must be much smaller than the smallest size of the opening in the mask and hence the device, that can be fabricated. The amount of light exposure required, the thickness of the photo resist, the time of exposure required are all purely empirical observations. Further, the semiconductor industry found several empirical procedures whereby devices of size comparable to the wavelength and in some cases slightly smaller have been fabricated. As mentioned earlier, the Raleigh criterion expects the smallest dimension to be large compared to the wavelength. The recent success has been quite surprising to many

physicists and engineers. There is however no real violation of the basic physics. The description of the fabrication process has only one purpose, to show that the technology required for fabricating these "quantum" devices is largely empirical and phenomenological.

VIII.5 Testing the MOSFET devices

In the case of the individual steps of fabrication such as diffusion of dopants and formation of the oxide, the experimental errors result is small variations. However the processes are defined by relations between independent (temperature, time) and dependent variables (dopant concentration, oxide thickness). The functional dependencies are experimentally verified and can be used for interpolation and extrapolation as required even if these are not normally derived from fundamental physics. Even this is not true for lithography where one takes a pass/fail criterion for any parameter. However, fabricating a typical MOSFET requires many such steps and the complex interrelationships and errors in all these processes means that only a statistical evaluation of the "functionality" of these devices can be made. Thus instead of phenomenological relations, purely statistical relationships enter the picture. A device manufacturer claims that the projected "functional" life time of the device is typically 15 years. Obviously there is no sense in actually testing the device for 15 years and even then there would be only a statistical result. Some fraction of the devices would fail very early and a small number may not fail even after many years beyond the claimed lifetime of 15 years. Subjecting such devices to much harsher conditions than are normally encountered, (higher temperatures, larger voltages etc) the manufacturer develops statistical failure models that enable him to expect that only very small fraction (that can be tolerated commercially) would fail before the stated life time. Once again this is a cursory look at what is an extremely complex test and evaluation process. However the key inference for us is the realization that the basic quantum and classical physics theories do not directly come in to these models. Indirectly they do influence both in the selection of the testing criteria and their expected influence on the device. For example, the expectation that harsher test conditions of higher temperature and voltage would increase the probability of failure is consistent with these theories but that is their only contribution.

VIII.6 A philosophical comment

The reason why this particular issue was discussed at some length is to highlight the absurdity of the calls by the philosophers and social scientists to reject physics as a model for their areas of activity. Actually this hierarchy of quantum physics, classical theories, phenomenological relations and empirical statistical evaluations is encountered in almost every area of physics and engineering. This is true even while discussing basic physics. One can at once see the similarity to the electrical resistance of metals discussed in an earlier chapter. The quantum theories can only give a functional dependence that can be expected for the variation of the resistance of a metal with temperature according to some simplified model of electrons in metals. As was pointed out this model fails during extrapolation and interpolation. If platinum is used as a thermometer and one starts testing its utility, a statistical rather than a phenomenological procedure has to be adopted.

The key conclusion is inescapable. Even in the context of technology developed using the basic understanding of physics and constrained by its theoretical structure, practical realization involves approximations and simplifications. Extensive use of the theoretical constructs of physics by philosophers a couple of hundred years ago was naturally counter-productive. At the same time it has to be kept in mind that deliberate downplaying of the role of physics and ignoring the constraints imposed by fundamental physics is also counterproductive. It has been pointed out that the basic structure of physics is based on a desire to provide an objective description of nature. It should be expected that these theories constrain the rest be they phenomenological descriptions in physics or commercialized technology. In the next chapter the constraints of physics on chemistry, biology and medicine are examined in some detail. This begins the core of our effort to answer the question of how well we know.

IX

HOW PHYSICS RELATES TO CHEMISTRY BIOLOGY AND MEDICINE

IX.1 The relationship between physics chemistry biology and medicine

The relationship between physics, chemistry, biology and medicine is very similar to the relationship between concept, design, fabrication and testing of the MOSFET device described in the last chapter. In each case, the basic quantum mechanical description of reality, the fundamental physics becomes progressively less relevant in practice while providing contingent limitations. Phenomenological, and quantitative descriptions are followed by purely statistical testing procedures as one proceeds to either testing of the MOSFET device or to medicine. The relationship between physics, chemistry, biology and medicine is philosophically controversial but this will be briefly discussed. More importantly, discussing the relationship between physics and medicine reveals one profound lesson that is not encountered when a MOSFET is discussed. This is the resistance to acceptance of the contingent limitations of physics. No matter how well we know something it may not be easy to accept it. As we move further into more complex issues in life, discussed in later chapters such situations are encountered more often.

Consider the burning of a small lump of coal, the simplest example of chemistry. At the atomic level this involves breaking the chemical bonds that bind atoms of carbon in coal and those that bond the atoms of oxygen in the molecules of air. This is followed by the formation of new bonds between carbon and oxygen atoms leading to the formation of molecules of carbon dioxide. The rearrangement of bonds between atoms is in the language of fundamental quantum mechanics equivalent to alteration in the electron densities in molecules. However when such reactions are studied, the issues of concern are the energy required to initiate and sustain the reaction, the kinetics of the reaction process and the role of possible catalysts that can alter the energy requirements and kinetics without being directly involved in the reaction. Burning of coal is an exothermic reaction, it results in the generation of heat. Thus once initiated the reaction will proceed to completion as long as carbon and oxygen atoms are available. The energy obtained by the formation of carbon dioxide can initiate the breaking of more carbon to carbon and oxygen to oxygen bonds. These atoms are then available to form more carbon to oxygen bonds or carbon dioxide molecules.

There are other reactions such as the one between citric acid and sodium bicarbonate which are endothermic. A continuous supply of energy from outside is required for the reaction to continue. A phenomenological study of the energy barriers and their alteration by catalysts is more important and fruitful than discussion of electron densities. It is in practice impossible to calculate the rate at which a reaction proceeds from the electron densities.

The oxidation of silicon discussed in the case of a MOSFET is another example of a chemical reaction. In device fabrication when a higher rate of formation of oxide is required, steam is used instead of oxygen. The choice is defined by empirical observation not by theoretical calculations. It is not possible to prove from physics that steam would increase the growth rates. At the same time the concept of chemical bonds is experimentally verifiable. When the formation of a bond between atoms is probed very carefully using femto second (one millionth of one billionth of a second) laser pulses, the alteration of the electronic clouds resulting in the formation of the bonds and the resultant energy barriers can all be observed. Thus there are

contingent limitations on chemical reactions and chemistry born out of the quantum physics of electron processes.

IX.2 Physics contingencies in biology

Since Lavoisier showed that oxygen is consumed and carbon dioxide is given off during respiration and that the heat produced by respiration was equal to the heat produced when the same amount of oxygen was used to burn charcoal there has never been any biochemical reaction that violated the rules of chemistry. Despite molecules being extremely complex, the reactions underlying all physiological processes are similar to other chemical reactions. Not long after Lavoisier, Wohler synthesized urea from inorganic compounds to demonstrate that biological molecules are made of the same atoms as the rest of the universe.

Physics provides contingent restrictions on the structure and functionality of biological molecules. A few examples illustrate this. A molecule of DNA, is shown in figure IX.1. The biological functioning of DNA is contingent on the separation of the two strands for copying the genetic information available on each strand. This separation without disturbing the individual strands is possible only because the bonding between the strands is through hydrogen bonds which are much weaker than the covalent bonds between the consecutive sugars on the individual strands. The chains are complimentary. When the nucleotide on one chain is Adenine(A),

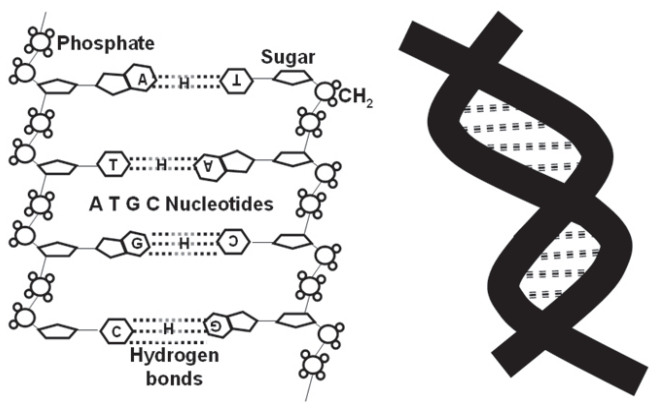

Figure IX.1 Structure of DNA

the one on the other is Thymine (T). If there is Cytosine(C) on one there is Guanine(G) on the other. This complimentarily is possible because three hydrogen bonds form between C and G while two form between A and T. In fact every investigation of the structure and functionality of bio-molecules displays similar evidence. In some cases such as hydrogen bonding between strands, it is difficult to even imagine alternate possibilities.

Not just the structure but even the kinetics of biochemical reactions displays similar contingencies of chemistry and physics. As an example we can consider the basic processes by which energy required for physiological processes is made available. In even the most simple single cell organism, there are thousands of biochemical reactions. At any given moment the rate of any required reaction may have to be to be increased or decreased in response to demands on the organism. The simple universal way in which this is accomplished is shown in figure IX.2. The molecule adenine tri-phosphate (ATP) is called the energy currency of the cell. It attaches one of the phosphate radicals to the required bio-molecule. This makes the molecule more energetic. Thus the energy barrier to the biochemical reaction is effectively lower and hence the speed of the reaction enhanced. Ultimately, the living organism converts the adenine di-phsophate (ADP) to ATP using the energy from sun light. Not only is this process elegant, it closely mirrors other chemical reactions and processes that can be investigated both qualitatively and quantitatively using standard methods of chemical investigation.

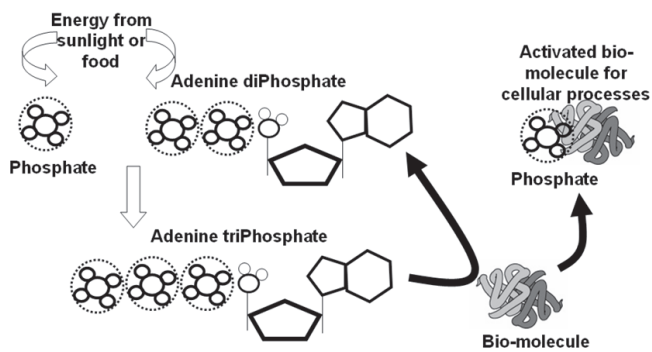

Figure IX.2 Schematic of the ADP-ATP reactions

IX.3 Homeostasis and the role of medicine

For any meaningful discussion of the relationship of physics as the underlying fundamental framework and medicine, it is necessary to recognize clearly the condition of homeostasis that is the essential feature of any living organism. All living beings, while maintaining the internal order, continuously exchange both matter and energy with the external surroundings. It is this maintenance of internal order that distinguishes a living organism from a chemical system. If one considers any living organism, the ingestion of food does not merely provide energy. Some of the atoms in the living organism are replaced by the atoms in the food. Thus the quip by Feynman. "What is this brain of mine? Last week's potatoes". The atoms which once constituted the potato now form part of the brain. The Greeks considered a logical challenge. They imagined a ship. Every part of the ship, the wood, the sails, the ropes and even the nails are slowly and over time replaced with new ones. The logical conundrum was to answer "when was the ship converted from the old to the new?" They never came up with a convincing answer. For the present discussion the logical and philosophical arguments are irrelevant. However the imagery is very useful. Every living organism enacts this process of change as long as it is alive.

Further, every cell, whether it is one among many in a patri dish or one among the many constituting a living animal is subject to changing external environmental conditions. The process described above, continuous renewal and maintaining the internal order has to be continued while responding to the challenges posed by an often hostile and changing environment. This ability is homeostasis. To accomplish this, the living organism employs many overlapping loops of feedback control. When the temperature rises, a human sweats to compensate for the unwanted heat flow from the environment. The evaporation of sweat causes cooling. At the same time, for the cells on the skin, the wetness is a change in the external environment for which they have to respond in order to maintain the homeostasis. There are some interesting philosophical issues concerning these feedback loops that will be discussed in a later chapter. For the present, it is important to note that when one talks of medicine as a means of assisting the living organism, one is referring to not merely a group

of atoms, but to the maintenance of dynamic order. Thus medicine or any other medical procedure has to support this self regulating mechanism, avoid suffering of animals and humans and prolong this self regulation to longer durations. Thus the contingent limitations on the regulatory mechanism due to physics and chemistry have to be identified without disturbing or stopping this regulator mechanism.

The key relevance for our discussion is that both disease and recovery are statistical in nature. Medicine is an area where only statistical analysis is possible. As is obvious this is similar to the case of the MOSFET. Even though a MOSFET is a ridiculously simple device in comparison to any living organism, the testing and prognosis of the device had to be performed by a statistical analysis. No wonder that as most doctors explain, medicine is "more an art than a science". The ability to restart this process of life is extremely limited even in the present advanced technological age. This places a huge burden on the practitioners of this "art" to deliver the greatest accuracy in the shortest time. Some of the key issues relevant in particular to humans will be discussed in chapter X.

IX.4 Medicine and scientific enquiry

As discussed above, a living system inherently has the capacity to maintain an internal order in the face of damage imposed on it by the environment. Thus the efficacy of any medical intervention has to be evaluated only in terms of expected change in the quality of life, extension of duration and reduction of suffering. This analysis has to be purely statistical. The most accepted scientific approach in medicine is a double blind test.

The first requirement to investigate the usefulness of a medicine in treating a medical condition is to find a group of people who suffer from this condition. All these, called the "test subjects" have to be at the same level of health. Thus one has to be reasonably sure that the course of the illness would be same without medical intervention. They are divided randomly into two groups. One group, the control group, receives a placebo, a substance that is known to be both harmless and useless in treating the medical condition. The other group, receives the actual drug or medical treatment. None of the

subjects knows whether he or she is in the control or test group. In a double blind test, the investigators also do not know who are in the control group. At the conclusion of the trial, the differences in the recovery of the two groups are determined. There has to be a statistically significant increase in the response of the test group to accept the medicine as effective for treating the medical condition.

This requirement of a placebo became standard procedure when it was realized that there are significant changes in human recovery purely due to the psychological belief of being treated. Even under the best of circumstances, the double blind test has severe limitations. These aspects will be further considered in chapter X. For the present we assume that statistical evidence based on such tests is the basis of medical science and inquire into the contingent limitations on these statistical interferences by fundamental scientific theories of physics.

IX.5 Physics based contingencies in medicine and resistance to their acceptance

The above discussion highlights the major problem in identifying the hierarchical limitations imposed on medical procedures by fundamental science. In the case of the famous example of relativistic correction to Newtonian physics, the limitations are clear even in the mathematics. The mathematics of relativity simplifies to classical Newtonian mechanics when the velocities of the objects are small compared to that of light, one of the objects is very massive compared to the other and the distance between the objects is large compared to their diameters. When chemical bonds and reactions are considered, while the study of chemistry is empirical and phenomenological, it has been possible to experimentally verify their conformity to the underlying fundamental physics. It is difficult to link even the basic theories of biology, the Linnaeus classification which relates all living organisms thorough a rank based classification or Darwinian evolution which provides a mechanism for the emergence of all living organisms related through such a classification to physics. Some indirect linkages between biology and physics exist and are considered in a later chapter on evolutionary explanations. It has been ridiculously easy to rule out the pre-Newtonian ideas of

physics. The very first simple experiments had disproved the physics of Aristotle or the phlogiston theory of burning. Questioning the efficacy of ancient medical systems has not been as easy.

Since all medicine is supported only by statistical evidence, it is easy to believe that trial and error over a long period might have provided viable medical treatments. This is the argument in support of traditional knowledge. There are several historical examples of traditional medicines that have been accepted and adopted by modern medicine. None is more famous than quinine. Extensive use of quinine had a major impact on human society. For example, it permitted large scale migration of Indians to Africa and the Caribbean during the nineteenth century.

Despite such examples however, applying the analysis presented in the previous chapters shows that such examples are improbable. Consider that a treatment "xyz" has been claimed in traditional medicine to be an effective treatment for a medical condition "abc". In contrast to the current practice of medical science this conclusion is based not on statistical evaluation and double blind trials but purely experience of traditional societies. Let us consider the possible methodology by which the traditional knowledge could have accumulated and compare this with the description of randomness and bias in earlier chapters. Consider that to validate a traditional medical practice a double blind test has been now performed and really confirms that the benefit attributed to the use of "xyz" is statistically significant. But because of our understanding of uncertainty and error we recognize that there are both successes (S) and failures (F) in individuals constituting the test and control groups. During the traditional societal history, the double blind trial was not used but individual trials of "xyz" should also have resulted in a mixture of successes (S) and failures (F). After all, there is no logical reason why the results in this scenario had to be different from the test group in the double blind trial. So the question of whether traditional societal experience could result in identification of "xyz" as a good medicine purely from trial and error is equivalent to asking if the preponderance of successes over failures can be identified from a sequence without statistical analysis. This was exactly the situation considered in chapter III with the series of coin tosses. Even if the

coin is heavily biased (0.66 was described), it is not possible to look at a series of finite length and without the use of statistical analysis guess if the coin is biased. In exactly the same way, even if we assume that the traditional medical procedure is effective in two out of every three trials, there can be a series of several failures. There were a series of consecutive tails even though the coin would come up heads twice in every three trials. On top of this, even when the recipe was recorded, knowledge in traditional societies did not include records of the successes and failures. So the traditional knowledge is actually the experience is of a single "traditional medical man" and a short sequence of trials.

Thus a logical examination of the process whereby the traditional knowledge could have accumulated, shows that it is improbable for experience to identify beneficial treatments since the relevant statistical analysis is not performed. While it appears to be logical to claim that traditional knowledge is distilled from experience and make the usual laudatory qualitative statements, chances of this occurring are pretty slim. This is the reason why traditional practices can be both good and bad. Just as in the quest for miracles and other areas of human endeavor that we will discuss in later chapters, selective memory (remembering only the successes) and emotional biases are responsible for most of the claims. Confidence in traditional medical knowledge is thus quite misplaced. Experience in the absence of statistics cannot act as a filter for viable procedures. Traditional medicine is only one among the various examples of alternate medicine which are based on "experience" and weak statistical evidence.

Some "alternate medical procedures" also violate the hierarchical or contingent constraints of physics. Homeopathy is the most famous example. As is well known homeopathy claims that decrease in concentration of the medicine increases its potency. This is directly contradictory to our understanding of biochemical processes. Molecular picture of materials suggests that at severe dilutions, even one molecule of the medicine may not be retained. The philosophical idea that a "memory" is retained even at such dilutions is simply incompatible with fundamental physics. The intermolecular and interatomic forces are well understood and unless

new forces of nature are hypothesized such memory effects are not possible. The problem is not in hypothesizing such forces, the problem is making them compatible with the rest of the observation of physics and chemistry. This is similar to the case of variable radio active life times discussed earlier. In the case of some other "alternate medical" approaches, it is not easy to identify such a contingent limitation. Acupuncture is one example among traditional medical systems. It is interesting to note that not withstanding extensive support for "alternate medicine", there are far more practitioners than investigators of these claims. There are casual statements like "why has everything to be proved according to the tenets of modern science". The real reason is obvious. There is a possibility of success as a practitioner. There is nothing to be gained by being an investigator and coming up with no positive results.

If restrictions based on current ideas of physics and chemistry are not accepted, disciplines such as medicine suffer from a problem of plenty. There can be too many possibilities. If one wants to treat a disease by herbs, the number of possible species available in any forest is mind boggling. The number of non-scientific approaches that can be tried is unlimited. Investigating each of these ideas and confirming their failure is limited by the availability of resources for performing these investigations.

As can be seen from the above discussion, it is not even necessary to evaluate traditional medicines and homeopathy with proper double blind investigations to be most skeptical about them. If a traditional medicine is really effective, quinine is the most famous example, even the most rudimentary statistical investigation will suffice. As Feynman pointed out, if a scientific investigation is on the correct track, when the experiments are repeated the evidence become more and more robust. In examples such as parapsychology, ideas that are contrary to science and fundamentally flawed, the evidence fails to improve with repetition. This is the case corresponding to small fluctuations in the observed ratio of heads and tails in an unbiased coin. Such arguments based on physics are not very palatable to society. This is not even confined to alternate medicine. The support for homeopathy is matched by the willingness of mainstream medical researchers to suspect and investigate the linkage between cancer and

power lines or use of mobiles. There is absolute unwillingness to accept that unless fundamental physics is completely re-written, it is impossible for the extremely low energy photons in power lines and mobiles to cause any change in the biochemical processes. The strength of the linkage, in line with Feynman's observation refuses to improve in survey after survey. Fluctuations that cause some surveys to show some "statistically significant" link are used to demand social action or at least more investigation. That a photon picture of electromagnetic fields is an integral part of fundamental physics, that Planck's law has been verified innumerable times, that the energy per photon is too small to alter the electron configurations in bio-molecules and that without those changes there cannot be any cancer is simply ignored. What remains is a mystic belief that such changes are possible in complex systems.

Another similar example is the fear of microscopic quantities of poisons, for example pesticides. Heavy metals are stored by the human body and will eventually cause major health problems. This is the reason why the mercury content in a fish is far higher than in the sea water. Even without evidence of any equivalent mechanism of similar retention of pesticides, there are demands for mandating unbelievably low levels of pesticides in aerated drinks. We discuss in the later chapters the role of economic prosperity in making ever stringent guarantees of safety and the support of other ideologies for such demands.

Thus, as one seeks contingent limitations of fundamental physics in chemistry, biology and medicine, one sees a progression. The contingent limitations are completely accepted in chemistry. A chemist is secure in the conviction that to perform good chemistry, phenomenological and empirical approaches are more useful than fundamental physics. At the same time there is complete agreement that when investigated, quantum process would limit chemistry and that such investigations are an integral part of chemistry. In the realm of biology, there is a justified insistence that the utility of physics contingency is quite limited. There is a philosophical argument that the most important aspects of biology science can never be reduced to physics and chemistry. It is claimed that, while structure and functionality of bio-molecules can be profitably studied, it is

impossible to break the organism as a whole into meaningful constituent parts that can be fruitfully investigated using physics or chemistry. However, even in mainstream medicine there is opposition to limiting investigations based on contingent limitations of fundamental physics.

———————

Summary

What Can Be Learnt From Physical Sciences

In the previous four chapters, a selective and extremely brief description of vast areas of human knowledge in physical sciences has been presented. Fundamental physics has a core mathematical structure that is absolutely necessary for any objective description of reality. Thus, ideas that contradict the basic fundamental structure can be summarily rejected. One example so rejected was the possibility that the radioactive lifetimes are not constant but variable. But as the selected examples discussed show, even in physics, it is not usual for the properties of complex entities to be constructed from fundamental forces. As P W Anderson stated "reductionist hypothesis does not imply constructionist one". Knowing fundamental laws of physics does not imply the ability to reconstruct all features of the universe based on them. Even in physics new concepts are hypothesized at every level of complexity. These do not contradict fundamental forces and symmetries but are not derivable form them. Instead, phenomenological ideas are generated for understanding experimental results. These are then linked to fundamental physics. Philosophically, the situation with chemistry or biology is no different. Hierarchy or no hierarchy, at each level of complexity understanding requires the same level of creativity and inspiration.

Another lesson from the study of physical sciences is the limitation of extrapolation and interpolation. Even when the physical observables are continuously variable and described by mathematical functional forms, extrapolation and interpolation are always provisional. Most often, failure of these provisional mathematical forms is recognized because of experimental evidence. It is rarely derived from fundamental physics. The only major exception is the relativistic correction. The fundamental symmetry, (the Poincare symmetry) has been the driving force behind the realization that constancy of mass is violated when approaching velocity of light.

The resistance to contingencies imposed by physics with respect to homeopathy and mobile mania illustrates the trouble with the effort to understand "how well we know". The answers to really complex problems of the society that will be taken up in the next part of this book are most often resisted. On one hand, the core strength of fundamental physics in not recognized and its aid is not accepted in weeding out approaches as with homeopathy and fear of mobile radiation. At the same time, statistical significance, a conclusion that is not at all robust is trusted for extrapolation without limit. Mathematics and physics show that careful use of the reductionist approach of science is required to assess how well we know the answers to complex questions. As with Gödel's proof in mathematics, despite strong human desire, approximate knowledge is all that is possible for many complex problems. The key issue once again is how approximate?

———————

Part Three

What Is The Science Behind An Ideology

"The time has come," the Walrus said,
"To talk of many things:
Of shoes—and ships—and sealing-wax—
Of cabbages—and kings—
And why the sea is boiling hot—
And whether pigs have wings."

Lewis Carroll

The discussion in parts I and II is only a cursory look at what is well known. At best a novel point of view can be claimed. Even the last bit about homeopathy and mobile mania is hardly new. Most sympathizers of homeopathy and mobile-cancer links concede that their views are contrary to physics. If pressed they resort to a justification based on the complexity of living systems. As one moves to the real huge problems associated with human society, novel prescriptions encounter strong resistance. It is no use claiming with T S Eliot that "I Tiresias, old man with wrinkled dugs. Perceived the scene, and foretold the rest—". To begin with the extremely modest accomplishments of the author of such sage advice will not be lost sight of. The aura of greatness in one discipline often permits learned

discourse in others. Rare is a Feynman who modestly admits that "outside his own discipline the expert is as ignorant as the next man".

As if this major drawback is not sufficient, the conclusions drawn herein will be rarely palatable to anyone. It is difficult to accept that a million workers in an established field are all running round and round in circles, that scholastic description does not have practical utility and as the child said in Anderson's fairytale, "the emperor is naked". Above all else, realism is often equated with pessimism. Empiricism has grown into a major branch of philosophy. However, converting empiricism to a philosophy has limited utility. There is something in the human psyche that is attracted by grand illusions of knowledge as power. However logical they might be, restrictions on human abilities are resented. While the idea of the golden mean has been appreciated since Socrates, it has been worshipped from far as an unrealistic ideal.

But fools venture where angels fear to tread. The next few chapters discuss issues such as global warming, economics, evolutionary psychology etc drawing heavily on the discussion in the first two parts. The primary message from science in the first two parts was caution against extrapolation. In every major problem of societal interest, positions are passionately and emotionally held. These positions are what are being termed ideologies in the present description. Analysis reveals the weak or nonexistent extrapolation from science that underlies not merely one among contesting ideologies but of all ideologies. The universality of fundamental physics is being imitated without recognizing the multi-tier linkages in science and engineering as illustrated by the example of the MOSFET. Most ideologies begin with statements that are supported by evidence of statistical significance or logic. They are then converted to commandments for societal action. Science as we saw is a curious mixture of ignorance and uncertainty. The goal of the next few chapters is to highlight the limited scope for extrapolation in finding solutions for these huge problems.

———————

X

MEDICINE:HEALTH AND SOCIETY

X.1 Medicine and ideology

Before trying to understand the science (if any) that lies behind any ideology, it is most appropriate to look at the issues of medicine, health and society. To be sure there is nothing in medicine that can be designated as an ideology. Support for either "scientific" or "alternate" medicine can be very passionate but it is not an ideology. At the same time, both for an individual and the society, health is the most cherished desire. It is a great problem for the patient, the doctor and the society to select the appropriate medical procedure. Therefore it is informative to ask if the selection of a specific medical intervention can be an obviously rational decision. The analysis is performed using a simple example that has been termed the "doctor's dilemma". This is not a dilemma of diagnosis by the doctor nor does it refer to the fear of a patient about the choice of the physician and acceptance of the diagnosis. George Bernard Shaw identified another doctor's dilemma, between his avowed desire to cure the patient and the financial loss he will suffer as a consequence of doing so. These are all real enough. While modern medical science has given a variety of tools to assist the doctor in the diagnosis, integrating all the information available

is a skill acquired with experience. The doctor's dilemma as discussed here is a philosophical tool to argue that utilization of a science does not depend only on the validity of the science. A study of medical matters is extremely useful as a paradigm for the analysis of other human desires which are sought to be accomplished by various ideologies. Fortunately, unlike in other areas, the conundrums and issues in medicine are accepted by all. Hopefully, the present analysis throws strong light on the problems taken up in succeeding chapters.

X.2 The doctor's dilemma

Historically, a medical practitioner was associated with religion. This reflected both the deep fear of disease and limited human capabilities. One feels, justifiably, that one has come a lot farther since Alexander the Great said "physicians from the entire world are collaborating while watching me die". But even today, the dilemma of a doctor is rarely understood or appreciated by the common man. The essential problem can be understood by considering the simple example outlined here. Assume that an individual is subject to periodic and severe headaches. Aspirin relieves the pain much faster than not using any medicine. In repeated trials, the duration and intensity of suffering is significantly less with the use of aspirin. The "statistical significance" has been established. However there is a complication. The patient is also susceptible to nausea and stomach irritation due to aspirin. For the sake of argument we assume that this has a much lower probability. Thus, this severe reaction occurs only on a few occasions following the use of aspirin. This can also be investigated using normal statistical analysis. For the present discussion it is assumed that this connection is also statistically significant. Now the doctor's dilemma is complete. If the doctor prescribes aspirin, there is a risk of suffering due to the reaction and if he does not he is not acting for reduction of suffering. In making this caricature of a medical issue, we are ignoring various real medical concerns such as the possibility that the headache is an indicator of some serious ailment. We are considering a very simple example so that the doctor's diagnostic abilities and practical constraints of time and sincerity can also be ignored. Any amount of further scientific research cannot resolve the dilemma. Only the confidence level of statistical significance can be altered.

As mentioned in an earlier chapter, statistical significance is accepted when the probability that the observed phenomenon is caused by randomness is less than 5%. Further research can reduce this probability for both the favorable (pain relief) and unfavorable (nausea) consequences. Newer medicines that replace aspirin or other treatment to reduce the nausea and other adverse reactions are possible. (in this case for example by prescription of an antacid). Progress of medical science is based on such approaches. But these do not logically prove the impossibility of similar dilemmas with their use. There is no "rational solution". The doctor's dilemma is ubiquitous and many a human problem shares both the conflict between "benefit" and "risk" and the absence of a rational resolution.

There is a famous story about Buridan's Donkey. This animal was supposed to be perfectly rational. Placed equidistant from two equal stacks of hay, being rational it could not decide which way to move. After all both stacks of hay are equidistant and they are also equally attractive. So the rational donkey did not find any argument supporting a movement towards one rather than the other. It thus starved to death. Doctors do not spend all their time wringing their hands in despair any more than a real donkey would starve to death when left between two heaps of hay. The reason for this resolution is apparent. A doctor does not look for such a "perfectly logical" answer.

A doctor may be influenced by his personal philosophy in his choice. His decision may be influenced by many psychological and emotional factors just as any human decision. However, he or she is very conscious of the fact that there is no "strictly rational" method to resolve this issue. This limitation has a superficial similarity to the hail stone numbers encountered earlier. When one process made the number larger and the other smaller, the net result was impossible to guess. Here, the benefits and risks cannot even be compared since they are not numbers which can be added or subtracted. One encounters a similar lack of a rational answer in both cases.

X.3 Medical progress and medicine as a science.

It is most important to emphasize that doctor's dilemma is merely a useful tool for the analysis in the present discussion and not

any kind of claim that medical research is not rational or that there is no progress in medicine. While the contingent limitations on "traditional and alternate medicines" have been highlighted in the previous chapter, success in prolonging life is the greatest achievement of humans. Over the years, medical science has helped in finding better cures for diseases and prolonging life. When you prolong one person's life even by one day you have accomplished in some sense far more than any other thing that a human being can achieve. After all, we have so far very limited success in restarting the complex life processes if they actually do stop.

Despite glorious accomplishments, the methodology of modern medical science has limitations. Medical research is based squarely on the double blind test. The limitations of this approach become apparent when we recapitulate the discussions in chapter IV and these are important for the discussion in the next few sections. The double blind test assumes that the members of the test and control groups are "identical". But no two human beings other than identical twins are really "identical". It not easy to select a group that will have an identical course of illness in the absence of any intervention. Even for simple diseases, a few will be cured very fast and a few will linger on for a long duration. There may be hidden variables, genetic or behavioral that may not be considered significant based on current knowledge. For example, before the linkage between cigarette smoking and health was established, smoking was not considered significant and not taken into consideration in selecting the groups for trials. Absence of such hidden interfering traits in any current trial cannot be guaranteed.

What does it really mean to say that aspirin provides relief to headache? Headache can be mild or severe. In addition to being subjective, it is often impossible to compare the intensity of the headache that the same individual has suffered on two different occasions. Ignoring this complexity for a minute it is possible to imagine a plot of the intensity of headache against time duration. One can in principle, draw two curves as in figure X.1, the first when aspirin was not used and the second when aspirin was taken as soon as headache was perceived. The curve shows the problem with most diseases. They are self limiting. Eventually the headache will subside naturally.

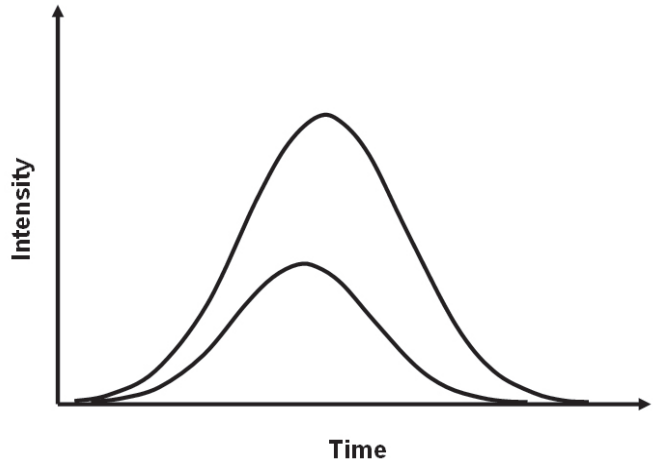

Figure X.1 Possible variation of intensity of headache with time with and without Aspirin

Repeating the effort we shall assume results in approximately similar curves and one gets two broad maxima. Obviously aspirin does significantly reduce both duration and intensity of headache. But in evaluating the medicine for statistical significance, all this complexity is ignored. The statistical significance simply confirms that medical intervention reduces the duration of suffering.

As discussed earlier when the differences between the two bell curves are not very large, there is significant overlap. In the present context this means some members of the control group have recovered faster than at least some members of the group subjected to treatment. For example, let the mean duration of life without treatment for a disease be "x" and with treatment "y". Research could show that the inequality $x<y$ is "statistically significant". However, as the example of the heights of men and women discussed in an earlier chapter showed, there is a probability for a random woman to be taller. Similarly, if the standard deviations are large enough, there is a probability for the life time of a random patient with treatment (y) being smaller that the life time of a random patient without treatment(x). Not every patient who has been provided the treatment has a benefit as compared to every single patient without the treatment. We noticed that if the mean values are 10 and 12 respectively and the standard deviation is 1, the probability would be 0.17. (See figures IV.5 and IV.6). If we assume that the statistics in the present case are

similar, in roughly in one case out of six, the treatment is counter productive. Even if other measurable quantities, such as blood sugar levels for a diabetic, blood pressure for hypertension etc., rather than subjective criterion like the intensity of headache are considered, medical research rarely enquires to see if the variables obey a Gaussian distribution or estimate the probabilities. In particular, the need for speed in identifying useful medical practices is enormous. Thus, the enormously more effort, time and money required for collecting the data and determining the probabilities may not be justified. Experimentally determining the probability does not in any case resolve the doctor's dilemma.

The medical fraternity is acutely aware of the limitations that exist in the science at their disposal. The limitations outlined above make the great progress in public health that have been actually achieved really a miracle. However, the problems to be discussed in the next few sections have to be understood in the context of the basic logical problem in the methodology of medicine outlined above.

X.4 The doctor's dilemma : Psychological, economic and societal dimensions

As medical science progresses, alternative procedures with fewer or less severe negative aspects are discovered. Aspirin has been largely replaced by other drugs like paracetamol for headache. So the doctor's dilemma becomes purely notional in this case. If instead of choosing an almost ridiculously silly example of headache, a major disease, cancer is chosen and the treatment is chemotherapy instead of aspirin, the dilemma at first appears to have vanished. The adverse reactions are simply accepted as a payment for the higher goal namely prolonging life. However, if the actual prolongation of life is quite small, balancing the benefit against the suffering imposed by the treatment is difficult. Medical progress enhances the human capabilities and makes the resolution of the dilemma practically simple only in some cases. That is not a rational solution, merely a practical resolution. This is in the same spirit as the practical resolution of the "induction" conundrum. The fact that sun has risen a thousand times in the past is no guarantee that it will do so tomorrow. But for practical purposes it is sensible to think that it will. In ancient Greece, unraveling

the Gordian knot was said to ensure becoming the king of the world. Alexander famously cut the knot. Every decision taken to resolve the doctor's dilemma is similar.

Consider a variant of the dilemma. The doctor is faced with a terminally ill patient with no acceptable medical treatment available. Now the benefit is non existent. But the knowledge about the placebo effect makes any physician pause. R K Narayan wrote an interesting story called "The Doctor's Word". The fictional doctor is normally completely rational and famous for announcing the sentence, "you will live" or " you will not". He finds himself unable to say the latter to a friend and announces "you will live". He is surprised to see the patient recover thanks to the belief in the word. This fiction resonates with our experience. Confidence in the doctor, in the placebo and a positive frame of mind have all been statistically established to contribute to recovery. In the story, it is the past reputation of the doctor that creates the confidence in the patient. The confidence that was generated by the "rational" (and honest?) statements of the past. So one sees another dilemma even in this fiction. Unless the doctor is honest to begin with he cannot produce miraculous cure through his word! At the same time, this is literature and not reality. No one in the real world would have that kind of trust in any doctor. But this unrealistic example like the simplified doctor's dilemma serves as a useful tool for thought. One can also bring in a "patient's dilemma" when two doctor's differ. The poor patient has to compare the capabilities of the two doctors. Socrates argued more than two thousand years ago that only one doctor can evaluate another. Consequently, selecting a doctor is not "rational".

The issues of alternate medicine now come into focus. When the foundation of an alternate medical science directly contradicts established science, it becomes quite problematic to support such procedures. This is true of homeopathy in the case of medical treatment and of the fear of electromagnetic radiation in the case of power line panic and mobile mania. In both cases there is absolutely no physical mechanism that can cause positive or negative changes in the biological processes. The latest effort at a scientific examination over thousands of patients has shown that the benefits of homeopathy are no better than a placebo. However, placebos and psychosomatic

disorders having physical symptoms originating from mental or emotional causes make the issue rather complicated. What if the "alternate non-scientific" treatment helps the patient because of these factors?

The fact that placebos sometimes work is clear evidence for the positive impact of psychology on health issues. The impact of stress and phobias is the corresponding evidence for negative impact. There has been some suggestion that part of the support for homeopathy is the tendency of the practitioners to catalogue all the symptoms and spend significant time with the patient rather than curtly ignoring symptoms that are considered irrelevant according to established medical science. This could enhance the "psychological self help". Sustained economic development may have indirectly helped the psychological desire for alternate medicine. Both when the ailment is not amenable to modern medicine (terminal cancer for example) or when the ailment is benign but not significantly controlled by medicines, (arthritic pain for example) the desire for alternate treatments is significant in economically advanced societies.

Many questions can be raised. Should the doctor buildup (false) hope? Should the doctor be held legally responsible for providing detailed description of the possible side effects? Is that practically possible? What would be the effect of such "disclosures" on patient psychology? Then one has endless complexities brought in by economic and societal factors. Even if the patient is "paying" for the medical services personally, the available resources are being locked up. Every hospital service that is being used by one patient is not available to others for that duration. Estimates show that in the United States, a patient spends up to half the life's savings on medical expenses for the last year of life. It is not possible to say if this is good or bad. There is no rational way of deciding between the requirements of two patients but such decisions are taken in real life. Military medical records routinely refer to drastic decisions taken in the battle field, to prefer those that "can be saved".

Similar "non scientific" criteria are employed to decide the place of a patient on a waiting list of recipients for donated organs when demand is more than availability. The criteria are not merely

arbitrary. They are all examples of the doctor's dilemma. The benefit to one patient causes a loss to another. There is no resolution only a human convention.

We can also consider the societal demand for universal vaccination. Vaccination for infectious diseases is not only a protection of the individual. It also reduces the chances of infecting others. This societal benefit is now to be balanced against any adverse reactions that have to be borne by the individual and not the society. It is simply non scientific to argue that the "vaccine is safe". There is always a finite possibility of adverse reactions. There is no such thing as absolute safety as will be discussed in a later section.

X.5 The doctor's dilemma and the prisoner's dilemma

It is worthwhile to contrast the doctor's dilemma as described earlier with the "prisoner's dilemma" which is very famous and has created a discipline of research, game theory. This dilemma involves two criminals who have together committed a crime. The police do not have a strong case so if the criminals cooperate with each other and remain silent, both will get a small prison sentence. On the other hand if one of them gives evidence against the other, he can escape punishment or get a very mild punishment. The other will of course get much more severe punishment. Human psychology being what it is, doubtful of the other persons intentions, usually both actually end up giving evidence against each other and both get the severe punishment. The dilemma is usually depicted in the form of a table (Table X.1) of possibilities called a payoff matrix.

Table X.1

Payoff matrix for the prisoner's dilemma

	Prisoner B Stays Silent	Prisoner B Betrays
Prisoner A Stays Silent	Each serves 6 months	Prisoner A:10 years Prisoner B:goes free
Prisoner A Betrays	Prisoner A: goes free Prisoner B: 10 years	Prisoner A: goes free

Extensive research literature has been created discussing more complex problems using the concept of a payoff matrix. The key point is the existence of a "rational" solution. The prisoner's dilemma becomes an interesting study of human psychology. The game theory and pay off matrices are all designed to "understand" human problems. The goal is to understand why rational answers are not preferred by humans and sometimes to predict possible outcomes given this basic understanding of human psychology. The doctor's dilemma on the other hand highlights a problem that has no rational solution. In contrast with the prisoner's dilemma, neither rational analysis nor more science can help. Identifying the point at which this condition, (the absence of a rational solution rather than merely the inability of humans to accept or implement a rational solution) emerges is the key to understanding how well we know to resolve large human problems. This brief introduction to game theories has been included here merely to point out that the dilemmas of life as discussed here are not amenable to game theory.

X.6 Medical progress and the precautionary principle

In an earlier chapter, the unwillingness to accept contingent physics based explanation of the fears of electromagnetic radiation was mentioned. Similar is the fear of extremely low concentrations of poisons, in particular pesticides. The reason for an unending demand for more investigations and warnings about such dangers is partly social psychology and partly economic progress. The risk perception in a society is related to the economic capabilities and there is a continuous up gradation of the minimum standards.

Unfortunately, when evaluating risks with such low probabilities, there are always statistical quirks that lead to false positives. It is relatively easy to perform a double blind test to test the toxicity of a pesticide and determine a safe limit. This usually refers to immediate adverse reaction. This may not be satisfactory. For example, cigarette smoking is injurious to health following long term use, not immediately. Further research may indicate long term problems that may emerge. However, it is extremely more difficult to confirm that regular ingestion through, for example food or water, of one thousandth of the designated safe limit will not have any

undesirable consequences. The problem was theoretically discussed in the case of base line fallacy. When the probability is low, the effective accuracy of test becomes quite low. It was pointed out that it is difficult to find a method for the confirmed detection of low probability events like earthquakes. If the probability is one per million, a test with a 99% accuracy for prediction will give so many false positives that it is practically useless. It is equally impossible to rule out the possibility of events whose probability is very low. As the number of trials increases, the confidence in the experimentally determined value of the bias increases. However if the bias is say 0.9999, determining this "accurately" is next to impossible. In the present case one is demanding a confirmation of low probability. If the probability of negative effects was not low it would have been easily identified. Thus it is never possible to "rule out" such a result. If sufficient number of studies are undertaken some false positives will always emerge. This statistical reasoning is applicable to concerns such as the influence of mobile phone radiation or low concentration poisons. There will always be a few positives which lead to demands at least for more studies if not immediate action. These negative observations are matched by the spontaneous miraculous survivals that are the experience of every experienced medical practitioner. These cases only offer encouragement to claims of both "miraculous divine intervention" by the religious minded and the "holistic, beyond conventional science alternate medical systems". It is no wonder that just as one cannot rationally rule out all fears, one cannot explain all survivals.

Aside from this desire for "no risk" solutions, medical issues cannot be evaluated without reference to the societal values and norms. Consider one of the biggest challenges to medical procedures, abortion. Even if the religious arguments are completely ignored, the success of the previous fifty years in increased survival of premature babies makes supporting the so called "right to abortion" quite a complex issue. If fetuses of 24 weeks can survive, is it a satisfactory decision to permit medical termination of pregnancy at 20 weeks? No rational answer exists and perhaps this quandary belongs more to societal issues than medicine. It is true that most arguments against abortions are the result of its conflict with Christian religious values. In India where there is no equivalent religious pressure, there is no

controversy about abortion itself. But this very different societal norm has resulted in the abhorrent practice of selective abortion of female fetuses. This has resulted in the abnormal sex ratios of nearly 800 female births against 1000 male births in many parts of India.

Another paradoxical example is the tendency of individuals to reject some established medical procedure. While an individual rejecting small pox vaccination in the nineteenth century was making a very risky decision, a refusal by certain groups not to avail blood transfusions in the current era has very small attendant risk. It is a moot point if this risk is larger than that of many adventure sports such as bungee jumping, sky diving or even motor cycle driving. The latest controversy regarding universal immunization for cervical cancer has several parameters. There is an economic issue. Public immunization will reduce costs and help commercial interests. The vaccine helps people with a particular life style (multiple sexual partners). Those with an alternate life style have very low risk of the disease and are not in need of the vaccination. The safe life style has religious sanction. That section of the society not only questions the public expenditure but also see the vaccination as an encouragement to a life style not supported by their religious views. However, apart from precisely determining the protection that the vaccine provides, science and more research cannot resolve the issue at all. For completeness, we have to admit that the utility of the vaccine is not universally accepted as significant. However, such research findings are countered with accusations of religious bias.

Ethical doubts will appear even with the cornerstone of medical research, the double blind trial. It is necessary and extremely useful. However, use of a placebo for control purposes appears to be less than fair treatment for the patients when a medicine for HIV is on trial. Are the patients not being given false hope? If the placebo group does so badly compared to the drug group is it ethical to deny the drug to the placebo group? But the "scientific validation" of the treatment is dependent on completing the trial. Every significant advance in medical intervention in human reproduction, (test tube babies), stem cell research and even human aging comes bundled with a host of ethical issues that are seriously debated and analyzed. From the current perspective, the role of philosophy and bioethics in

such cases also has to be recognized as a "fundamentally non-scientific approach". Bioethics cannot provide any "concrete" resolution in most cases. Very much like Feynman's comment about teaching being "ineffective in most cases and superfluous in the rest", the few simplified cases resolved by philosophy do not need the assistance. The real world problems are intractable with bioethical analysis as much as they are unsolved by mathematical rigor. Bioethics and philosophy are supposed to "sensitize" the doctors and the patients to taking the correct decisions. Similar is the hope of the religious that observance of the tenets leads to a morally correct judgment but through a process that cannot be analyzed! This can be called a heretical thought but as we shall see in a later chapter, such statements emerge once again in social sciences.

The label "doctor's dilemma" may be new or original. That these issues have no explicitly rational solution has been known since the time of Socrates. Generally it is expected that principles of utilitarianism (greatest good for the greatest number) or the Hippocratic principle (first of all do no harm) would guide the physician. But they do not offer any step by step analysis leading to the QED type of a proof of a theorem in geometry. It is quite clear that "rational" is being used here in a very restricted sense of being reducible to a series of small logical steps which cannot be individually refuted very much like theorem in geometry. As pointed out above, principles of bioethics and philosophy help in a way that is distinctly different from teaching a theorem in geometry.

One advantage with medical matters is that everyone readily admits that the problems are complex and that there are uncertain consequences of any intervention. Another is a general acceptance of the marginal benefits that accrue from advanced medical techniques. There is little doubt in anyone's mind that the vaccine for cervical cancer will not have the same impact on human health as the original vaccination for smallpox. Many doctors attribute two thirds of the improvement in longevity and general health over the last century to emphasis on hygiene and availability of better food. As we move to future chapters we encounter more complex problems associated with society at large. However, there is less willingness to accept such limitations. We will endeavor to understand those issues in the light

of the present analysis and seek to emphasize their similarity and the necessity for similar resolution.

Medicine with all the societal complications actually serves one individual. Even here the inability of science to provide rational resolution of issues is quite obvious. The fundamental issue is that nature of data obtained in medical research is incompatible with extrapolation and interpolation that constitutes science. It is this limitation in the nature of data that is at the crux of the search for an answer to the question of how well we know many other areas to be discussed in the following chapters.

———————

XI
ENVIRONMENT:SCIENCE AND ACTION

XI.1 Scientific measurement of human impact on the environment

Environmental issues dominate current human discussions. Human impact on the environment and the sustainability of the current lifestyle are issues of societal importance which are passionately and emotionally discussed. A brief analysis of the science behind the environmental issues will be taken up concentrating on the strength of the science. One is asking the usual question, how well we know the various statements that are being made. As in the previous chapter, the brief analysis is used to evaluate if conclusions drawn would be contingent on further research and study. If problems have no rational solution, increased mathematical, experimental or analytical rigor can only provide intellectual satisfaction. Environmentalism is an ideology in the sense in which this word is being used here (without any value judgment). The goal is to evaluate the strength of the science and the confidence in the suggested means of mitigating the human impact on the environment. While other environmental issues such as pollution, radioactivity, acid rain and ozone hole have been prominent societal concerns of the recent past, current interest is centered on

global warming due to human activity. This is the only issue being discussed here.

The current environmental concern is the continuous increase in the concentration of carbon dioxide in the atmosphere. The data most often quoted, reflecting the increase in the atmospheric concentration is presented in figure XI.1. The measurements depicted in the first curve are data since 1960, obtained at one place, the Mauna Loa observatory in Hawaii at an altitude of about 3000m above sea level.

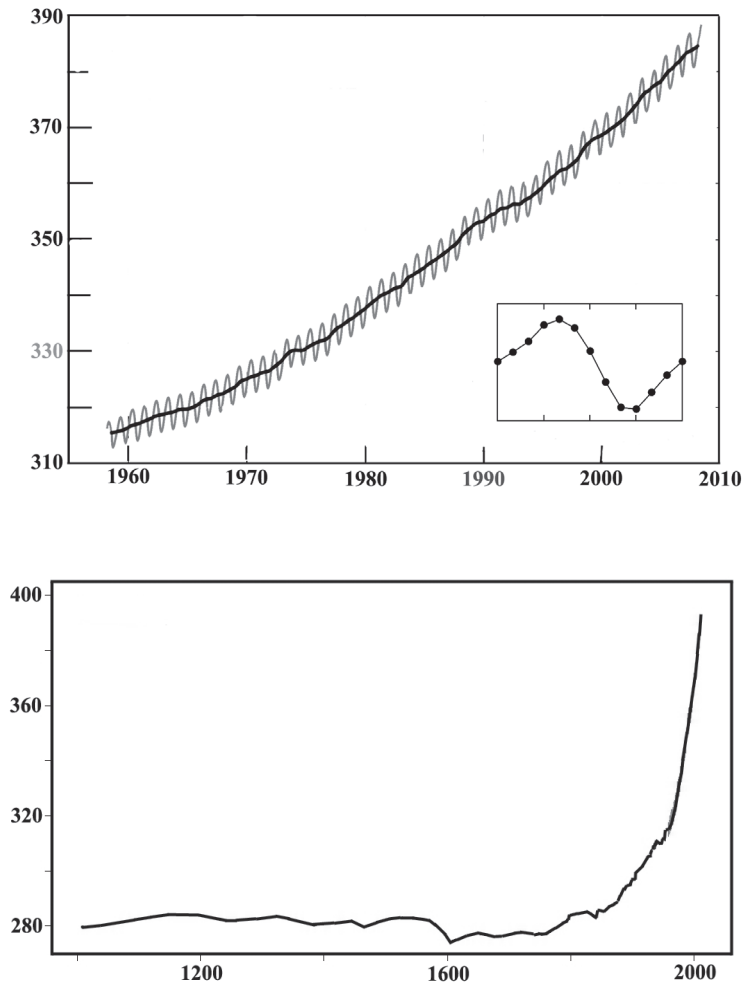

Figure XI.1 The Keeling curve showing atmospheric carbon dioxide concentration along with earlier data from ice cores

This is named the Keeling instrumental record, after the scientist who initiated the measurements. The second plot includes data from about 1000AD, mostly obtained from ice cores.

In both Greenland and Antarctica, the snow that has fallen in one year forms a distinct layer. These layers include gases trapped from the atmosphere. Layers have accumulated in both areas for thousands of years. These cores are carefully drilled out and the concentration of carbon dioxide in the trapped gas determined. The ice core data stretching back to several hundred thousand years are now available and show a fluctuating concentration of carbon dioxide varying between 200 to 300 ppm. The data from Mauna Loa, for the year 2010 is nearly 386 ppm, well above the historical figure justifying the assertion that this increase is caused by the burning of ever larger quantities of fossil fuels. The basic question that is always to be asked is the same, namely how well do we know that? It is most important to know if the concentration of carbon dioxide in the atmosphere can be measured accurately and reliably and if this can be attributed to fossil fuels.

The exhaust of an automobile would have from 5 to 15% by volume of carbon dioxide. If one were to make the silly mistake of measuring near the exhaust of a fire extinguisher in operation, one can even measure a 100% value. The concentration also changes with time. An average value can be calculated from data obtained at a given location. However it is important to consider if measurements such as those at Mauna Loa can be justified as average values for the whole of the earth?

If one considers another constituent of the atmosphere, water vapor, one notices very large changes. The relative humidity, the number of water vapor molecules in the air does change from time to time and place to place. At a few places in the Atacama Desert it has not rained in 400 years. The humidity never reaches 100%. In Mount Waialeale, Hawaii, it typically rains for 350 days in a year so typically the humidity reaches 100% everyday. So a global average value has no real meaning. This is an example of non-Gaussian distribution mentioned in an earlier chapter. Mean, median and mode can be calculated but impart no information.

However, fundamental physics of the molecules of water and of carbon dioxide is drastically different. A physicist would call the water molecule polar and of carbon dioxide non polar. Their melting and boiling points are vastly different. These are 0 and 100C for water at atmospheric pressure. Melting temperature of solid CO_2 is -78C and boiling temperature of liquid CO_2 is -57C at atmospheric pressure. As a consequence, at the temperatures normally encountered on the surface of earth, exchange of water molecules from the liquid and solid phases to the gas phase and vice versa are energetically possible. This physics is behind the various weather phenomena like rain, dew, snow or just plain variation of humidity. These continuous transformations from one phase of water to the other are extremely complex resulting in the extreme differences mentioned above between Atacama and Hawaii. An additional factor is the availability of plenty of water. If all the ice available in the world would melt, the sea level would increase by about 60m. So the amount of water available on the planet is really very large.

The possibility of such continuous exchange between phases of carbon dioxide is non-existent. Even if the local temperature at some point on the earth were to reach the required temperatures, the available quantities are too small. In the absence of these phase changes, the molecules of carbon dioxide are free particles that diffuse very fast. In the absence of any force, diffusion reduces concentration differences. The time taken for diffusion is quite small. Typical diffusion coefficients of molecules in the gas phase are 0.1-0.2 cm^2/s. To get feel of these numbers, the typical values for ink in water is about 0.000001 cm^2/s. A drop of ink completely disappears in a few seconds in a bucket of water. Note that the diffusion coefficients in gases are higher by several orders of magnitude. Thus any local change will be rapidly averaged out.

Carbon dioxide is heavier than air. But even this does not invalidate the calculation of an average carbon dioxide concentration. Even if one considers a totally static atmosphere, without wind, the molecules of carbon dioxide do not settle to the bottom. If a layer of carbon dioxide is introduced artificially on the floor of a room, for example from a fire extinguisher, it will stay near the floor only for a short while. The molecules of air and carbon dioxide will mix rapidly

because the mixed state is more disordered. Such a state has a lower energy and is always preferred. Energy is not needed to disorder the mixture of air and carbon dioxide. It is needed to form a separate lower layer of carbon dioxide. In the absence of wind, gravity would have caused the ratio of carbon dioxide to be higher at the lower altitudes. But such changes will be quite small (typically a few %). Winds will further reduce this gradient. Hydrogen and helium are lost from the atmosphere not because the hydrogen molecules rise to the higher altitudes like balloons. Hydrogen molecules rise to the upper atmosphere due to mixing before being lost due to their higher velocities in the upper atmosphere. The mean distance traveled by a molecule of air at the earth's surface, before colliding with another and changing direction is about one micron. Gases in the lower atmosphere are extremely well mixed and no changes in the mixing ratios of gases (other than water vapor) are observed till a height of about 100km. Since pressure, temperature etc are changing, the concentration of gas is determined as parts per million by volume. This is also called the mixing ratio. This number defines how abundant a gas like carbon dioxide is with respect to others. Gravity cannot cause any significant change in mixing ratio of carbon dioxide and phase changes responsible for changes in water vapor concentration do not contribute.

The local concentration of carbon dioxide on earth changes because of the carbon cycle. Plants absorb carbon dioxide to synthesize carbohydrates as every school book of science describes. All living organisms including plants produce carbon dioxide as a waste product of respiration. Among other processes that can locally change (increase or decrease) the concentration are human burning of biological products like wood or fossil fuels, decay of dead organisms, absorption into the surface of water and weathering of silicate rocks. The magnitude of these contributions and their change with time has to be taken into account before identifying the measured carbon dioxide concentration as a reasonable average value for the earth.

Once carbon dioxide enters the atmosphere, energy is required to remove it. This is due to the total entropy or disorder. The only natural process that can have the required energy is photosynthesis but even this is extremely inefficient. The energy in the carbohydrate

molecule is a tiny fraction of the solar energy that falls on the tree, typically less than 1%. Partly this is due to the low concentration of carbon dioxide. As molecules of air enter the leaves, there is no filter to selectively permit only carbon dioxide to enter. Most molecules that enter are nitrogen which cannot be used by the leaf. So the process of consumption of carbon dioxide is very slow. Weathering of silicate rocks, another known process of carbon dioxide consumption is also a slow process since carbon dioxide in the gas phase has to dissolve in rain water, an inefficient process.

There is only one significant physical process for absorption of carbon dioxide namely absorption of the gas on water surface. This is possible because the surface of the earth is largely water. Carbon dioxide dissolves in the sea water (most surface water is sea water) when the temperature is low and is released back into the atmosphere when it warms up. Thus the need to keep carbonated drinks refrigerated to preserve their fizz or dissolved carbon dioxide.

While photosynthesis and sea water absorption processes appear very complicated, the data presented already provides an interesting conclusion regarding their magnitude. The data collected in Mauna Loa show a cyclic variation that is shown in the inset. There is a maximum concentration during early winter of the northern hemisphere and a minimum during the spring-summer. A look at the globe will convince one that the ratio of land to ocean is about 1 to 1.5 in the northern hemisphere and in the southern hemisphere it is 1 to 4. During the spring in northern hemisphere, plant growth is a maximum while the seas in the southern hemisphere are coolest. Both contribute to more absorption of carbon dioxide. During the winter in the northern hemisphere these are reversed. The observed oscillatory behavior do not specify the magnitude of the carbon dioxide that these processes add or remove from the atmosphere. There can be extremely large exchanges to and from the atmosphere. However the presence of the observed oscillations demonstrates that these always take place at much shorter time scales.

Incidentally, it should be kept in mind that every measurement is actually an average. When a simple mercury thermometer is used to measure temperature, it does not respond to rapid changes. For

example a fan can be rotated at a few revolutions per minute and allowed to cut the sunrays falling on the mercury thermometer. It could be expected that the temperature will be lower when the blades of the fan block the rays of the sun. However, the mercury thermometer will only read some average value. Such changes can only be measured with some specially designed thermometers. Averaging is inherent to every measurement. There is no difference between using the physical body (mercury and glass) to perform the averaging or making measurements using a fast thermometer and performing the averaging mathematically. This is the principle of signal recovery by averaging that was discussed in an earlier chapter. The oscillatory behavior of the carbon dioxide concentration shows that large changes due to photosynthesis and ocean water absorption/desorption occur at short time scales.

Another key observation is the magnitude of these changes. On an average, the size of these cyclic changes is less than the observed increase in about five years. It is almost impossible to imagine any other large scale source (apart from anthropogenic) that could be responsible for the observed magnitude of the rate at which carbon dioxide concentration is increasing. Note that the observatory is near the equator and there are no strongly defined seasons for either growth of plants or sea temperature. So clearly these changes are not of local origin.

Further, one can estimate the quantities of fossil fuels and other bio-materials burnt by humans. The increase in carbon dioxide accounts for about 40% of the yearly human contribution. This immediately shows that the observed increase is of human origin. If the increase were from other natural sources such as the release from the sea due to warming of oceans for example, the quantity should have been larger than the human contribution. A higher concentration in the atmosphere increases the probability of absorption in the ocean surface in a simple physical process. Thus the entire human contribution would not remain in the atmosphere as the 40% figure shows. Since the human consumption of fossil fuels is definitely increasing, the limited increase in the atmosphere points to the apparent increase in the carrying capacity of the planet.

It is now clear that the first question in the issue of environment can be answered with confidence. The carbon dioxide concentration measured at Mauna Loa is clearly a valid average value for the concentration in the earth's atmosphere and the increase clearly reflects human contribution. It is not a big surprise however. Ancient hunters had caused extinction of large animals and fishing has depleted the apparently limitless resources of the sea. There are limits to the capability to what can be absorbed by the environment and human activity is large enough to cross these limits.

XI. 2 Physics of greenhouse warming

Even conceding the above, since the concentration of carbon dioxide is quite low, it is necessary to justify that such low concentrations are important contributors to green house warming of the earth. Labeling the increase in the earth's temperature due to the presence of certain gases in the environment as greenhouse effect is actually misleading. A greenhouse has a transparent roof that permits the entry of solar radiation. In such a building, convection is prevented and the higher temperature is mostly the result of trapping the hot air inside the building. An automobile left in the sun, another popular illustration of greenhouse phenomenon becomes unbearably hot only when the windows are closed and circulation of air is prevented. However the name is too common for change. Greenhouse warming in the earth's atmosphere is caused by the absorption and emission of radiation by molecules. There is greenhouse warming on the earth even without any human contribution.

The temperature of the earth becomes a stable value when the amount of radiation it receives from the sun is equal to the amount it radiates. The same principle is observed when an object is placed near a fire. Its temperature will increase to higher values till the energy gained from the fire and lost to surroundings become equal and the body attains the temperature appropriate to that distance. If the object is moved nearer or further, its temperature will change to a new value after a time delay. In general the temperature will also depend on how well the body absorbs heat. A black body out in the sun is hotter than a white body. This property is called emissivity. A perfectly black body has an emissivity value 1 and a perfect reflector 0. Most real

bodies will have some value in between. As a first approximation, the contribution to the heat balance on the earth from other sources has been neglected. This is the correct scientific approach since quantitatively contributions from sources such as volcanoes or geothermal heat, radioactive decay in the earth's core or tidal friction etc are smaller than the contribution of the sun.

The radiation emitted by any body is called black body radiation and is known to be a universal curve shown in figure XI.2. In the figure, the intensity of radiation is plotted as a function of the wavelength for the universe at large, the earth and the sun. The universality of this curve is confirmed by the experimentally determined radiation from the universe at large. All radiation consists of photons or particles. At very large wavelengths, the total energy emitted is very small since the photons of these wavelengths carry very little energy. At very short wavelengths the energy per photon is very high but the body itself emits only few photons so total the energy lost at these wavelengths is once again small. Thus each body emits a maximum amount of energy at a specific intermediate wavelength. This peak wavelength depends on the temperature.

The hot body like the sun has a peak in the visible or ultraviolet regions. Light in these wavelengths has much higher energies while the universe, which acts as a black body at a temperature of -271C, emits in the microwave regions. The sun acts as a body with a temperature of ~5,000C. The earth, at an intermediate temperature of about 14C emits in the infrared region. The total energy that is radiated per unit area is determined by a formula called the Stefan-Boltzman law. The amount of radiation depends on the fourth power of absolute

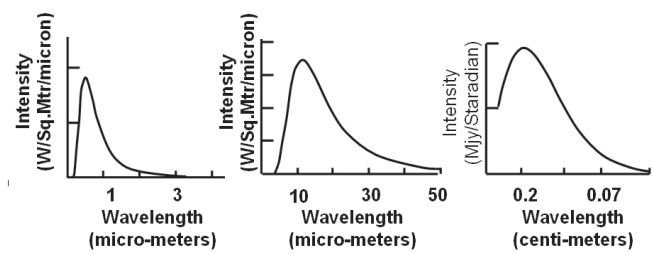

Figure XI.2 The black body radiation emitted by the universe, the earth and the sun

temperature (temperature in C + 273). Thus a body will radiate about 3.5 times more energy if it is at the temperature of boiling water (100C) that it would if it were at the temperature of melting ice (0C). It is simple arithmetic to calculate the temperature of the earth required to radiate the entire energy it receives from the sun. For the earth to maintain equilibrium, the temperature of the earth would have to be 6C. This assumes that the entire solar radiation is incident on the earth.

In the case of the earth, it is known from experimental measurements that approximately 30% of the radiation of the sun is reflected back into space. This so called albedo of the earth is obviously an approximate average value since the earth is not a smooth spherical ball of uniform color. Typically fresh snow reflects about 80% of incident radiation and water or a green forest cover will reflect 10%. If the albedo of 0.3 is assumed, the temperature would be -18C instead of 6C for equilibrium. In either case the temperature is lower than what one perceives as the average temperature of the earth. Large areas of the earth, near the equator are never as cold as even 6C. The lower temperatures of winter in Northern hemisphere are balanced by the warm summer in the south. It is unrealistic to expect that there is no reflection. After all the earth can be photographed from the space and that is impossible unless there is some reflection. The average temperature is experimentally about 14C.

This argument considers only the current conditions on earth. Given the sun's radiation at the earth's orbit, an experimentally observed quantity, any object in earth's orbit must have much lower temperature than is observed. Only one physical mechanism can explain the observed higher temperature. There must be a trapping of the amount of heat leaving the earth. Gases in the atmosphere can absorb the infrared radiation emitted by the earth. But any molecule that absorbs radiation will emit it almost instantaneously. However, while the molecule is absorbing a photon from one direction, the surface of the earth as the photon is going into space, it emits a photon in any direction. Some of these radiated photons, usually at even longer wavelengths can then be absorbed by other molecules in the atmosphere or the solid earth. If the emission in the longer wavelengths where the Earth radiates is reduced more than the reflection of the

shorter wavelengths of the sun's radiation due to the albedo phenomenon, the average temperature of the earth can be higher than the simple estimate calculated assuming that the earth is a black body.

Oxygen, nitrogen or argon, the common gases in the atmosphere are incapable of absorbing the infrared radiation. The prominent gases that can accomplish this are water vapor, carbon dioxide and methane. All three are produced by natural processes and exist in the atmosphere without human contribution. These are called the greenhouse gases. Since experimentally, average temperature of the earth is higher than the black body estimate, greenhouse effect caused by these gases is operating. It is a bit more complicated to estimate quantitatively the relative importance or contribution of each of these gases. One thing is quite clear. The concentration of these gases in the atmosphere is quite low. Clearly even these are sufficient to cause significant increase in the average temperature.

Water vapor, the most abundant of the gases that can contribute to greenhouse warming is less than 2-3% of the earth's atmosphere. As was mentioned earlier, the variation of water vapor concentration is quite large and unlike carbon dioxide it does not appear to be meaningful to calculate an average the concentration value. For example, at 30C, if the concentration increases beyond 3% at sea level, there would be condensation of liquid water, limiting the concentration. Increase in temperature due to the other two gases, carbon dioxide and methane would lead to increased presence of water vapor leading to further increase in greenhouse warming. Clearly, there is a degree of complexity here. All three gases can contribute to the warming. However, whatever warming is caused by the other two gases will contribute more water vapor and consequently further warming. Since carbon dioxide dissolved in sea water exists, warming could also lead to further release of carbon dioxide from the sea. Some of the transformations, formation of clouds or snow fall can alter the albedo of the earth. Due to the formation of dew, rain, fog etc., a water molecule is estimated to stay in the atmosphere only for 7-10 days. This is another important parameter that defines how much greenhouse warming can be caused by each molecule of the gas added to the atmosphere. Methane concentrations are several orders of

magnitude lower than that of carbon dioxide. Its oxidation to carbon dioxide in the presence of UV radiation in the upper atmosphere is possible. Carbon dioxide due to its very low chemical reactivity and low boiling temperatures is quite stable and is estimated to stay in the atmosphere up to 100 years. Carbon dioxide concentrations in the ice cores have been determined dating back to nearly a million years. The data shows cyclic variations which will be discussed after the primary issue of determining the average temperatures is analyzed. For the present, it can be confirmed that low concentrations of these greenhouse gases can increase the average temperature of earth. Thus the possibility of an extra warming by the increased concentration of carbon dioxide in the atmosphere is certainly possible. However, the first requirement is to experimentally measure an authentic average temperature of the earth and confirm that it has increased over the recent years. One has to know how well the temperature and carbon dioxide concentration are correlated before trying to know if the correlation implies causation.

XI.3 Measuring the average temperature of earth

From simple perception or local measurement, one perceives large fluctuations in temperature between day and night and several other long term variations, the most important being the seasons. While it is generally true that the night is cooler than the day, exceptions, typically due to rain or dust storms are part of human experience. Even within the seasonal variations, heat waves and cold waves create large fluctuations over a period of several days or a couple of weeks. While it is still possible to guess that the average temperature in New Delhi is higher than in Siberia, it is not at all clear that the average temperature of the tropical city of New Delhi is cooler or hotter than an equatorial city like Colombo. Thus to look at the long term trend in the average temperature of a given location itself appears a difficult task not to speak of comparing the trend between two different locations.

Reasonably accurate records of long term temperatures are available only since about 1850 at selected urban locations. The most common data is the daily record of maximum and minimum temperatures. These temperatures in addition to the daily and season

variations reflect local conditions. A cloud cover in the local area could reduce the maximum temperature in summer while a clear night could lower the minimum temperature in winter. It is however unlikely that the same day in the calendar will have similar local phenomena. So the average temperature for a given day of the year (maximum+minimum)/2 could be determined from the available records. Clearly this is a simple arithmetic calculation and does not make much physical sense. There is no way the average as defined by (maximum +minimum)/2 is the correct temperature. It is not the most common temperature (a mode) nor is it guaranteed that the temperature is lower than this for half the day (a median). However, by averaging the mean temperature over several years for a given day one can hope to avoid local weather phenomena like rain, cloud or fog if they are unseasonal. So the average would reflect the seasonal variations or long term effects. But even these cannot be compared from place to place. Differences in altitude, latitude and proximity to the sea (among other factors) will create permanent temperature differences.

So the deviations from the average value over many years for each location are separately computed. Now these deviations from different regions can be compared. If all the deviations show the same trend, it is a significant observation. Deviations measured using different techniques, for example surface instruments and satellites can be compared. Once again similar trend of deviations increases confidence in the data. The deviations may reflect temporary changes due to a global phenomenon. For example a huge volcanic explosion can increase the reflection (albedo) and cause temporary global cooling. The series of these deviations are then averaged over both shorter and longer periods (quarterly, yearly or 5 yearly etc). As averaging is performed over longer periods, the influence of random events such as volcanic eruptions and even some of the long term phenomena like the recurring El-Nino are suppressed. Thus one obtains a residual trend. It is this trend that is held up as an evidence of global warming. Thus the confidence in the average deviation displayed in figure XI.3 is much higher and more physically significant than an average absolute temperature of the earth. One advantage of these deviations is the possibility of averaging these deviations observed from various locations or even various methods of analysis etc. The experimentally observed deviation of about 0.4C over 25

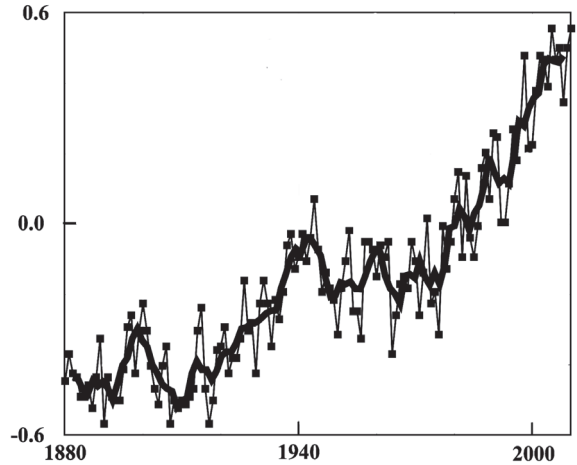

Figure XI.3 Recent average temperature deviation on earth

years becomes statistically significant only when the averaging is done over periods longer than 25-30 years. In popular discussions, this deviation is referred to as a "hockey stick" due to the visual similarity of the deviation in the graph with a nearly vertical hockey stick. The curve can be made to look like the infamous hockey stick by the simple expedient of plotting over a longer period. To include data from earlier periods, proxy measurements (temperatures obtained from other types of measurements) are included. The longer the period used, the better the visual of the hockey stick and unfortunately more partisan the arguments. But the short term data that is certainly more reliable than these extrapolations does indicate increasing deviation.

Overall, the empirical evidence confirming increasing deviation from the average is weaker than the corresponding evidence for concentration of carbon dioxide discussed earlier. This is best seen in the contrast between the regular features of the carbon dioxide variation as compared to the more noisy appearance of the variations in temperature. The support for an increase in temperature comes more from the variety of measurements and locations that confirm the warming collectively. Particularly, when the recent measurements are examined over a 25-30 year range, the increase is regularly substantiated. Thus one can on the basis of preponderance of evidence answer the question asked at the start of this section. We know pretty well

that the average temperature of the earth shows a small long term increase of about 0.5C over the past 25-30 years.

XI.4 Variation of temperature, GHG concentration and solar radiation

Clearly both the amount of solar energy incident on the earth and the concentration of greenhouse gases influence the eventual temperature of the earth. A detailed picture determined from ice core data for about half a million years of earth's history is provided by the data presented in figure XI.4. The first and third curves from the top are the carbon dioxide and methane concentrations which are directly measured from the gas trapped in the ice. The fourth is the concentration ratio of isotopes of $^{18}O/^{16}O$. The concentration of ^{18}O decreases with temperature. So the ratio indicates a relative change in temperature which is also plotted as the second curve. The correlation between the temperature and concentrations of methane and carbon dioxide is quite apparent. Small deviations between the two, particularly when the data are rapidly changing are obvious but the visually observable correlation is still quite impressive. The extremely complex temperature profile that one observes gives a first inclination that the relationship is not very simple. Since this data does not include any human contribution, the most natural and simple

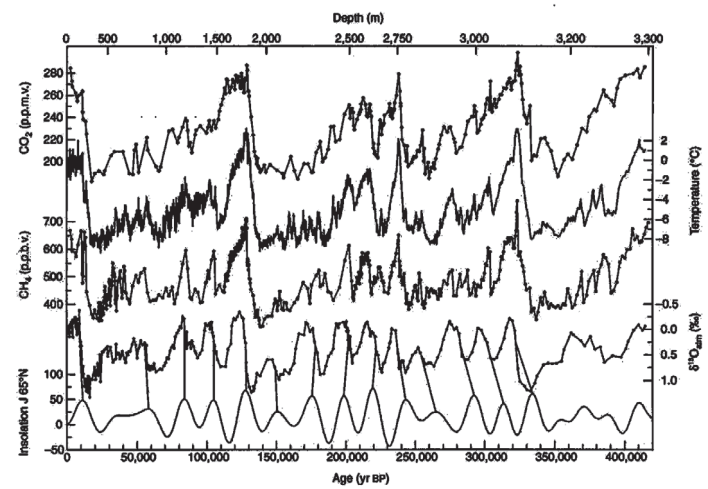

Figure XI.4 The data from ice cores

explanation would be that as and when the temperature increases due to increased availability of solar energy, the carbon dioxide concentration increases, primarily because of the release of the gas from the oceans.

However the estimated variation of the available solar energy plotted in the bottom of the plot for the 65N latitude does not visually show the same level of correlation with the temperatures and concentrations. Straight lines linking the available solar energy and the abrupt changes in the oxygen isotope ratio have been shown in the figure. While there are some instances when the two match, in many places these lines are not even vertical. This visual lack of similarity is actually quite significant. It shows straight away that the correlation between the temperature and green house gas concentrations is very complex. In addition to the prominent feature of the temperature with a period of 100,000 years, there are smaller significant features in the ice core data that do not have any apparent periodicity.

The theoretically calculated changes in solar energy availability on the earth at a specific latitude represent minor changes in earth's

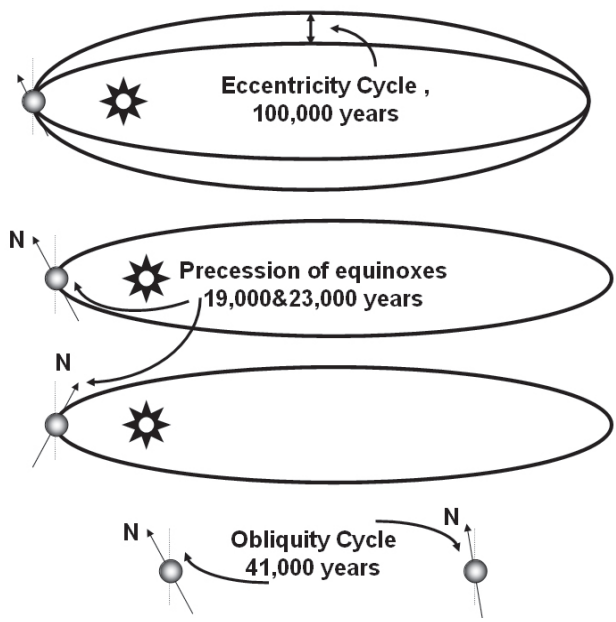

Figure XI.5 Variations in the earth's orbit with time

orbit around the sun. In addition to the seasonal changes in the availability of solar energy on the earth, there are several minor changes that can be deduced from experimental measurement of the orbit. Thus while the radiation available itself is not experimental data, the calculation is based on other experimental observations. This curve represents three different contributions that have been superimposed. The first is the variation of the eccentricity which varies from 0-0.07 with periods of 100 thousand years and 410 thousand years. The eccentricity is a measure of the extent to which the ellipse is deformed as shown in figure XI.5 at the top. All the changes shown in the figure are highly exaggerated.

The second is the season of the northern hemisphere when the earth is at perihelion. Currently the North Pole points towards the sun when the earth is at the perihelion (closest to the sun). This however varies with a period of approximately twenty thousand years. In the middle of this cycle, the North Pole points away from the sun at the perihelion. Finally the angle between the spin axis and the axis of orbital motion varies between 22^{o} and 24.5^{o} with a periodicity of 41 thousand years. These are independent contributions that alter the orbit of the earth from an ideal ellipse. The perturbations are small and more importantly independent of one another permitting them to be separated from experimental measurements. The deviations from the ideal behavior are caused by the gravitational forces due to other planets in the solar system, due to the earth being oblate, (the diameter at the equator being larger than the distance between the poles), due the friction between the atmosphere and solid earth etc. Some of these, like the influence of a planet like Venus or Jupiter on the orbit can be calculated in principle from basic gravitational law. This can be accomplished in practice because the force due to the sun is very much larger. So it is possible to treat these planetary influences as perturbations. Some like the contributions from the oblate shape or the losses in the atmosphere cannot be calculated from basic physics theory but have to be determined by phenomenological approaches. But even these are small and more importantly independent. This enables the clear separation of these perturbations.

The solar insolation as it is also called, is the most important variable that can be imagined simply because that defines the net

energy available. This changes with the latitude. The average insolation on the earth if calculated would not actually vary significantly. The correlation between changes at 65N latitude and the ice ages reflects geography, the larger landmass in the northern hemisphere. Even this correlation named the Milankovitch cycle reveals a subtle relationship between temperature and solar energy availability. The dominant 100000 year cycle in temperature does not correspond to the most significant cyclic variation of solar energy available even at 65N.

Thus what the cyclic variation in temperature actually shows is the importance of feedback mechanisms and nonlinear effects and correlated triggers. The carbon dioxide moisture relationship was already mentioned. Increased moisture can lead to cloud formation and increased albedo while lower temperatures can cause snow fall and increase in albedo. Another complication is the observed correlations between short term weather phenomena and the sun spot activity. Different mechanisms have been proposed for this though the total solar insolation varies only by 1 part in 1600 with the sunspot cycle.

All these complications confirm that the relationship between temperature and solar energy availability cannot be simply decomposed into dominant effect and perturbations. Statements like "it is CO_2" ,"it is not CO_2" and "it is all solar variation" etc are all too simplistic. Questions of whether CO_2 concentrations are "cause" or "consequence" are meaningless and unanswerable. If the various components of the system were not interrelated there would not have been the complex temperature record. Complex modeling will have to remain the only possible means of relating the temperature and green house gas concentrations and any effort at simplified explanations is futile.

The increase in temperature has been teased out of the available data with extreme difficulty. As mentioned earlier this evidence is not as robust as the increase in carbon dioxide concentration nor is the relation between the carbon dioxide concentration and the temperature increase as straightforward as attributing the increased carbon dioxide concentrations to human activity. The observed increase in carbon dioxide concentration due to human causes is large

enough to cause the approximately 0.5C increase in average temperature over the last 25-30 years. It is clearly illogical to claim that the observed level of carbon dioxide increases would have no impact on the climate. The rate of increase in anthropogenic carbon dioxide is also significantly larger than the rate of increase of various other natural phenomena that are correlated with weather and climate.

The variation of green house gas composition and temperature records over half a million years gives another indication. Large scale changes in climate can be triggered by many combinations of the various variables. Thus, one has certainly to worry about the carbon dioxide generated by human use. While the visual correlation in figure XI.4 is quite satisfactory, it appears very illogical to even consider modeling such a complex behavior with any degree of accuracy. No useful model can provide as an output such a complex curve. Clearly the situation with respect to climate models is similar to the doctor's dilemma discussed in the earlier chapter. A problem exists but its resolution cannot be proved rationally. While further refinement of the complex models is an ongoing activity that appeals to scientific curiosity and intellectual challenge, increased model accuracy by itself cannot resolve the dilemma of decision making. At the same time, many of the current passionate arguments about environment are based on details provided by the models. Some of the important warnings or fears will be considered individually in the next two sections to assess how well we know each of these. We then look at rational responses to the issue of global warming.

XI.5 Climate and weather

Three potentially problematic consequences of climate change caused by anthropogenic carbon dioxide have been highlighted and are often the focus for arguments in favor of immediate environmental action. These are increase in sea levels, changes in rainfall patterns and increased occurrence of severe weather phenomena. The common mantra of environmentalism is "weather is not climate" implying that the inability to predict short term weather is not related to confidence in climate change. Scientific prediction of local weather over anything more than a couple of days is no better than common place ideas of hot summers and cold winters. Consequently jokes about the

weatherman are most popular. Unfortunately, two of the three items listed above are weather related phenomena. The question of how well we know the consequences of climate change is thus more closely related to weather than the mantra allows.

Let us consider the experimental evidence of the claim that the global warming will lead to rise in sea level and consequent submersion of coastal areas. To begin, is a change in sea level experimentally measurable? While one expects the surface of water to be flat, on a global scale this is far from true. Changes of the order of 160m due to non uniform gravitational field of the earth have been mapped. Subsistence of land is another problem in accurately determining variation of the sea level at any location. The experimentally determined sea level variations vary significantly with location. The sea level rise over the past 6,000 years is estimated to be about 3m without any clear corresponding increase in temperature.

The best estimate of the current rate of increase is approximately 2mm/year based on terrestrial measurements and about 3 mm/year from satellite based measurements. These have to be considered in the context of the limitations mentioned in the earlier paragraph. As discussed earlier, a global temperature anomaly could be demonstrated. However, experimental evidence for a similar anomaly in sea levels aside from the long term raise mentioned earlier is not strong. It is quite possible to determine an apparent local increase in sea level but in the absence of a global trend, linkage to global warming is tenuous. While the available ice on continents if melted could raise sea levels by 60 meters, there is simply no consensus even among the proponents of environmental action of the actual magnitude of the increase in sea level. Some estimates of possible increases due to global warming are less than 2 meters in the worst case scenario and a most probable value of 0.5 m by 2100. However, there is a tendency to over emphasize the magnitude of the increase in sea levels in support of activism. Arguments about the loss of land space are emotionally appealing but simply overlook population increase. Most countries have more than doubled their population since 1960 and the resultant loss of per capita land availability is much larger than that attributed to sea level changes. How well do we know that sea level changes are real and related to global warming?

The evidence is weak relative to that for the temperature deviation. The claimed deviation of a few mm per year in the in sea level, correcting for tidal changes, erosion of coastal areas etc has not been confirmed to the same level of accuracy. It is possible that over-emphasis based on this weak evidence is in the long run counter-productive for rational handling of the environmental issues.

The second major catastrophic consequence of global climate change is expected to be change in rain fall patterns. The typical claim would be that "an increase in annual mean precipitation in high latitudes and Southeast Asia, and decreases in central Asia, the area around the Mediterranean, southern Africa and Australia is very likely. Higher greenhouse gases result in higher atmospheric temperatures and thus climate extremes will increase substantially". This is obviously a comparison between the pictures that emerge from the climate models with and without the greenhouse gas contributions. While lots of claims about actual determination of changes in local rainfall patterns correlated to the global warming are made, these are extremely unreliable. It is not at all clear that there is actually anything identified as a rainfall pattern even without the additional complication of its change. If there was one there would be no lack of short term prediction of weather. It is simply illogical to claim that the short term predictability (weather) is not possible and turn around and claim credibility for climate related changes in it. Currently as everyone knows reliable weather prediction is restricted to early warning based on radar observations for weather phenomena. The rest is more a convenient topic for amusing conversation.

The reason is quite obvious. While the temperature cycles, both daily and seasonal have at least a local average, there is no such average for rainfall or wind velocity and direction. It is the presence of these local averages that enabled the detection of a long term trend in average temperature. It is indeed possible to get an average value for the rainfall for a given day. Newspapers routinely mention the average rainfall for any given day. As usual the question is not if an average can be calculated but if it makes any physical sense to do so. Precipitation and wind velocity changes over short term are extreme. Rain could be a drizzle, a shower or a downpour and often all three occur within an hour. Local occurrence of cloud or storm can alter

the temperature but by much smaller amount. Temperature may change by 10-20% but not by orders of magnitude as is the case with rainfall. It was the deviations in temperature that could be compared from place to place and then provide a long term trend. Before considering the issue of rainfall one has to ensure that the long term averages have some kind of experimental consistency. The Indian experience in predicting monsoons is a lesson. Prediction of a normal monsoon is followed by observation of droughts and floods in local regions.

Anecdotal evidence of change in "weather" over a few decades is overwhelming. Virtually every individual comes forward with reports of such observation. The problem is to separate various possible contributions. In addition to climate change, one has possibilities of local changes due to urbanization, deforestation and other changes in land use. When we examined the claims of traditional medicine in an earlier chapter, we explored the unreliability of observations that have not been recorded, quantified and subjected to statistical analysis. Similarly, acceptance of the predicted rainfall pattern changes has to be preceded by analysis of experimental observations demonstrating changes similar to the long term heating.

The third and final class of major catastrophic predictions center on severe weather phenomena such as cyclones or hurricanes. There are claims that increased severity and frequency have already been observed. The statistical evidence for this is most suspect for several reasons. The classification of these major weather disturbances is at best qualitative. Criteria adopted for categorizing a hurricane as category III or IV etc are all very subjective. The number of such events is so small that identifying any statistical correlation is practically impossible. Rarity is a characteristic these severe disturbances share with earthquakes. The issues of base line fallacy that limit the statistical evidence of such events have already been discussed. While severe storms are more common than severe earthquakes, they are more complex and there is nothing similar to even the "Richter scale" for comparing two different hurricanes. The evaluation of the severity is clearly influenced by the damage that occurs. Unfortunately the damage in terms of human deaths and monetary losses are extremely dependent on the state of the economy. An earthquake of equal intensity would cause a thousand times more

fatalities but much smaller economic damage in an underdeveloped country such as Haiti in comparison to the USA. The damage also depends on the location. A severe storm in an uninhabited area of the earth would cause much less "damage". The reasons for the increased concern in these severe disturbances are the same as in the case of medical issues. Increase in population and wealth is leading to more support for the precautionary principle.

The above description is not a claim that the observed increase in carbon dioxide concentrations and the increase in temperature are harmless. It is not a denunciation of the climate models that predict increased in frequency of severe weather disturbances, changes in rainfall patterns and raise in sea levels. It is a simple caution that while the intellectual desire to model these disturbances is commendable, the confidence in human ability to confirm changes in rainfall patterns and frequency of storms is inherently very low. The sea level changes are more likely to be observable. These limitations on the human capabilities have to be accepted along with the warnings from the models and the extremely reliable estimates carbon dioxide concentration and a somewhat less reliable increase in global temperatures to decide on the human response to the issues.

XI.6 Complexity, non linearity, chaos and living planet

The next set of major concerns regarding climate change accept climate as an extremely complex system, influenced both by living organisms and the forces of inanimate nature. The first of these is an extension of the concerns regarding rain fall patterns that emerged from climate models. This is the concern that carbon dioxide concentrations and human influence of the climate may reach a tipping point leading to extremely rapid large scale changes. The second is the influence of rapid climatic changes on the other living organisms that are adapted by Darwinian evolution to a stable or slowly changing environment. The third view considers the earth, the atmosphere and the living beings as a single complex entity that is often compared to a single living organism called the Gaia. Thus there is a fear that Gaia itself maybe in death throes. We will now evaluate how well we know these concerns.

That climate and weather are complex is obvious to the most casual observer. While it is relatively easy to characterize winter as cold and summer as hot, a mild summer night could be cooler than a bright winter day. If we include rainfall and wind velocity, no clear pattern is easily visible. The data presented earlier of the temperatures inferred from ice core data is enough to demonstrate the complexity. Many of the physical mechanisms described earlier show that the relationship between various components is non linear. As one example, the relationships between the amount of water vapor in the atmosphere, formation of clouds and the resultant change in albedo are clearly not linear. This is distinct from the function being not linear as in the Stefan Boltzman law. In accordance with this law, the amount of energy radiated by earth varies as T^4 where T is the temperature in Kelvin. Thus a 10% change in temperature results in more than 10% increase in radiant energy. Conversely a 10% increase of solar energy causes much smaller increase in the temperature. Similarly the solar insolation varies inversely as the square of the distance from the sun. The non-linear relationships more significant for climate change models are not such simple functions from basic physics. These are similar to the recursive models that were encountered earlier in the case of deterministic chaos. Quite often the power relationships have non integral powers. It was pointed out during the discussion on least square fits that in such cases the fits and minima are not theoretically justified. In addition there are the complex iterative loops. An increase in temperature leads to an increase in water vapor content. This further enhances trapping of radiation by greenhouse effect and increases temperature. This causes a further increase in water vapor. This will result in formation of clouds reducing solar insolation and in case the humidity crosses the saturated vapor pressure there would be rain. There are numerous such links in the case of climate leading to complexity.

The existence of complex nonlinear linkages is very well known. In most cases it is nothing but simple physics. The key additional importance of the tipping point is the identification of one or more limits beyond which the consequences of climate change would become irreversible or highly accelerated. In most simple models of nonlinear systems, a small change in a variable would lead to major changes. Usually these changes are an indicator of a chaotic

system. The easiest example is the model of the deterministic chaos that was discussed in chapter II while describing repetitive mathematics. The deterministic chaos generated by the simple repetitive function $x(n) = r \times x(n-1) \times (1-x(n-1))$ was shown in figure II.3 and described in detail. Both unpredictability and self similarity could be observed in such a simple system. Research over the last few decades has shown that emergence of such patterns and chaos in models of nonlinear systems is extremely common. This easy possibility of chaotic response to minute changes is reflected in the oft quoted statement "fluttering of the butterfly's wings can cause a storm". That is mathematically true but obviously useless as a guide for human action. After all how well do we know that an earlier trigger has already not set in motion a chaotic consequence? And what is the relevance of the billions of flutters that have so far taken place?

Most descriptions of the consequences of tipping point in climate models do not follow this simple deterministic mathematical description. However it is important to remember that in non linear systems observation of such run away effects is very common. This reduces the confidence in the predictions of calamity following defined tipping points. In some sense a tipping point is almost inevitable in any model with complex interactions, and is not a surprising conclusion. The second major problem with these tipping points is their values being always just a bit higher than the current values of carbon dioxide emissions. This reminds one of the great example of Millikan's oil drop experiment cited by Feynman in his lecture on "cargo cult science". He relates, "When they got a number that was too high above Millikan's, they thought something must be wrong– and they would look for and find a reason why something might be wrong. When they got a number closer to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off,…." . In the current context, models with tipping points far into the future (2000ppm) or past (320ppm) would not be news worthy or even publication worthy!

The most obvious problem is the difficulty in pinning down climate change in the first place. Finding the rate of change of an imprecisely defined quantity is simply unscientific. Once again there is really no reason to discard the concept of a tipping point describing

a change in climate from one equilibrium state to another. The ice core data show that such tipping actually does occur. Each sharp change in the curve in figure XI.4 is an example of a tipping point. That is the simplest way to describe the observed fluctuations. However, when such fluctuations are observed all over the range of 200-300ppm carbon dioxide concentrations, support for a specific tipping point as a precursor for severe climatic changes is weak.

The second issue is the impact of rapid climate changes on living organisms adapted by Darwinian evolution to the current environment. This issue is more severe for the larger living organisms. For the unicellular organisms, adaptation is very fast as can be observed with the development of drug resistant species of germs. Speed of adaptation of even the larger animals must have been reasonable to accommodate the fluctuations recorded in the ice core data. More importantly, evidence for extinction of species due to human presence is widespread. The emergence of modern man about a hundred thousand years ago coincided with the extinction of a variety of large animals on several continents. The rapidly increasing human population has undoubtedly restricted the availability of habitat and resulted in extinction of species on the land and commercial fishing has contributed to severe reduction in populations in the sea. The recent observation of the large scale threat to frog species across the globe is well documented even if the cause is not very clear at present. Thus ascribing a major role for climate change in disappearance of species is suspect since other forms of human impact are far more important. In any case the tiger has become the "ward" of the humans irrespective of whether it lives in a zoo or a "natural habitat". Its survival is contingent on human support with or without climate change. There are current reports of extremely large losses in living organisms for example plankton in the sea. If the cause for such changes is really climate change as is claimed, there is more cause for resignation to the inevitable than motivation for remedial action. The current change in climate is small. Given the life of carbon dioxide in the atmosphere, even an immediate stopping of production of carbon dioxide by humans (an impossibility) would still result in two to three times the current change. Clearly if things are so delicate and there have been such drastic changes, there is little prospect of success. Given the resilience of life as we know it, it is more correct to attribute

the observed impact on organisms to loss of habitat and pollution than to climate change.

The belief in extremely sensitive and complex interrelationships between the physical earth, atmosphere, water and living organisms finds its ultimate expression in the Gaia hypothesis. This approach looks at the whole of earth and all the living organisms as constituting a single living entity named Gaia. The feedback loops linking various parts of the living and nonliving environment in this view constitute a homeostatic system maintaining on the planet geochemical and climatic conditions favorable for life.

While there are stabilizing feedback loops that have managed the earth's climatic conditions to be suited for some form of life or the other, there is certainly no centrally organized sense and correction. The change from a purely reducing environment of the early earth to the current oxygen rich environment is a consequence of living organisms. This has resulted in the change from the prokaryotes to the more advanced eukaryotes and finally the multi-cellular living organisms. Consider decay of living organisms. Increase in availability of dead bodies will increase organisms responsible for decay. Obviously this increase cannot continue indefinitely and will be self limiting. But this does not constitute any feedback process with a centralized control. The ability of a human body to reduce sugar levels though increase in insulin levels has limits but there is also a clearly distinguishable sensing and response mechanism. This centralized sense and correct mechanism is necessary for labeling an entity as a living organisms. We will return to explore the idea of the sense and response mechanisms in a later chapter.

The general deference to the idea of emergent properties, a more scientifically acceptable form of the concept of "irreducible complexity" provides a backdrop for the claim that homeostatic control on a planetary scale has been scientifically established. Such claims fail to admit that integrating this with established Darwinian evolution is difficult. One is reminded of "group selection", the idea that Darwinian evolution can operate at the level of groups. This is ardently supported despite the conflict with conventional Darwinian mechanism.

For the present, the relevant issue is the consequence for global climate change and human response. Unfortunately, the very generic, philosophical and sometime mystic descriptions permit a range of responses that are mutually contradictory. It is possible to demand the most stringent restriction of human contribution. It is also possible to claim that the homeostatic control makes pollution control unnecessary. It is possible to claim that space probes carrying life to other planets are the means of reproduction of Gaia. Humans will thus simultaneously become the sperms, the means of reproductive capacity and the cancer that is destroying the living planet. It has to be admitted that this imagery of sperm and cancer has not been employed earlier. The attractiveness of the Gaia hypothesis apart from the scientific merits is its resonance with the current human effort to consciously expand his brotherhood. Man's progress has been defined by his successive acceptance of the clan, the nation, the race and now the entire human population as part of his brotherhood. PETA (People for the Ethical Treatment of Animals) and Gaia are evidence of the latest expansion of this universal brotherhood. While one admits the animals, the other accepts the entire earth into the brotherhood.

The proponents of Gaia share with Malthus, rational apocalyptic fears of unprecedented devastation and ultimate doom. Applying rational arguments about growth of food and population, Malthus provided an early rational and scientific reflection of Christian religious feelings of the impending apocalypse. He logically argued that food production will increase in small steps, while the population is proportional to the number of people available and will increase exponentially. As was seen in figure VI.2, the exponential increase will soon overtake the linear increase. Thus inevitably (or so he argued) population will outrun food availability leading to starvation and death to a large fraction of population. Despite failures of such predictions, this fear of an apocalypse, continues to be repeated 150 years later by many, most notably by Ehrlich. The original proponent of Gaia, Lovelock has his own version in recent years.

A critical examination of the various potential catastrophes attributed to Global warming and climate change brings one to the conclusion that none of the fears can be ruled out on logical grounds as was possible with homeopathy or mobile mania in the earlier chapter

on medicine. However, the doctor' dilemma becomes a rational environmentalist's dilemma. It is not possible to provide any scientifically justified probabilities for the various catastrophic scenarios and this situation will not change with more research. Determining the probabilities and creating possible scenario's for projection is very common consequence of overusing mathematics. Feynman's analysis of the probabilities of failure of the space shuttle would be a good lesson for those claiming to provide such estimates. Inherently these probabilities are simply not determinable. Nevertheless, this overuse of mathematics enables one to do the easiest thing namely to fool oneself. Fortunately, Malthus comes to the rescue. The truly scientific contribution from Malthus was his influence on the development of Darwin's theory. On a smaller scale and to a more modest degree, the failure of the catastrophe predicted by him offers a good framework for analyzing the rational choices for environmental action.

XI.7 Rational possibilities for environmental action

The conclusion of the effort to find how well we know various environmental concerns ends in a dilemma. On one hand the enviroskeptics are dead wrong when they question the increase in carbon dioxide concentrations or human use of fossil fuels as its source. The increase in global temperatures is on balance of evidence quite real. Certainly the recent increase in concentration of carbon dioxide is large enough to account for the increase in temperature. The concentrations, much higher than in the ice core records can disturb the climate in unpredictable ways. On the other hand, the fears regarding changes in rainfall patterns or other major catastrophes cannot be scientifically verified because predicting the future evolution of a complex system and of low probability events is beyond human capability. There is no rigorous and unambiguous way to assess the actual probabilities of these possible catastrophic consequences. Thus once again there is the requirement for action, societal action in this case, without a "strictly rational" solution.

The story of the Buridan's Donkey described in the previous chapter, is a warning against ignoring the problem. The society better not starve to death waiting for a rational answer. While probabilities

for various environmental catastrophes cannot be scientifically provided, the catastrophes are certainly possible. Consequently, there is a case for concern. So an attempt at a practically rational action has to be taken up, just like the resolution of medical problems, as was discussed in the previous chapter.

The increase in carbon dioxide concentrations has one positive consequence to the society. The increase in carbon dioxide concentration is a consequence of the increased physical comfort of humans. Currently, energy usage and consequently the carbon dioxide emissions increase with "human development" whatever be the social and physical parameters that are used to characterize "human development". The United Nations Human Development Index increases rapidly as the per capita energy demand increases from a small value common in the underdeveloped world. The population of the world in 2011 is nearly 7 billion. One sixth of these are enjoying a way of life that the rest clearly find most attractive.

Clearly technological ability to deliver similar "human development" at significantly lower energy consumption or at least with drastically lower carbon dioxide emissions is not merely attractive; it is mandatory. Any call for stopping the rest of the world from "aping" the energy guzzlers will naturally fall on deaf ears. The guzzlers themselves cannot be easily shamed, frightened or forced into abandoning the use of energy when that usage clearly provides a more desirable life style.

Given that the earth is a finite world, there has to be a natural maximum sustainable population of humans. Whether this limit is small or large compared to current human population, optimizing the use of natural resources is a trivially correct approach. Similarly given the possibility of climate catastrophe even while the probabilities of such occurrences and the corresponding carbon dioxide levels remain unknown, technological effort to minimize the concentration is a trivially correct approach. The key question however is the relevance of the various possible proposals for improvements and a realistic estimate of the reduction in carbon dioxide emissions consequent to their implementation. We now consider two popular examples proposed as response to global warming.

The first of these examples is the use of solar cells to generate electricity without carbon dioxide emissions. There will be physics based limits on any carbonless energy technology and the solar cell is no exception. For example, the efficiency of a common single crystalline solar cell cannot exceed 30% due to basic physics reasons. Sunlight has photons of various energies. The solar cell will not absorb and use photons with energy smaller than some limit. Also the cell will ideally generate as output electrical energy equivalent to this limit even though some of the absorbed photons have much higher energies. Thus the cell looses some photons totally and a fraction of the energy from those that are actually absorbed. Both these limit the ideal theoretical efficiency. The best practical silicon device made in a laboratory has an efficiency of about 25%. Industrial production has not yet reached 20% and the cost of production is too high to allow this to be preferred means of generating electrical power for a home. Clearly improving the efficiency to the theoretical limit is a scientific and technological challenge. There are a number of social, political and economic actions that can also be proposed to enhance the use of solar cells. These could range from subsidies to the manufacturers, increasing taxes on petroleum products, social advocacy and supporting the research efforts.

The one advantage of the example of solar cells is the strength of the science. In the other example considered here for technological mitigation of climate change problems, even this is not well established. There are several proposals for geo-engineering. The claim being that the global problem of emissions should be corrected by a global engineering program. Injecting sulfates into the atmosphere, artificially brightening the clouds or using orbiting mirrors are proposed to increase the amount of sunlight the Earth reflects. Fertilization of the ocean with iron to promote algae formation is proposed to absorb carbon dioxide in the atmosphere. The more conventional proposals such as carbon capture and storage at the thermal power stations and increasing forest covers are not geo-engineering proposals.

If the standard question is now asked, “how well do we know that these measures will work?”, the answer is far from obvious. Scientific and technological progress cannot be guaranteed by research

funds. This is no different from the earlier statement that 2500 years have not proved whether there are an infinite number of twin primes. The efficacy of economic approaches cannot be different from their utility in other areas. Whatever be the rational arguments behind economic measures for mitigating climate change, the inherent limits of economic logic particularly on large economic programs will remain. Strength of the science of economics and other social sciences will be scientifically examined in succeeding chapters.

Climate is shown to be a complex system. If there is no real way of estimating the consequence of increase in carbon dioxide concentrations, it is equally difficult to accurately predict the consequences a geo-engineering project. Consequences of increasing the albedo of the planet by artificially enhancing the whiteness of clouds have to be estimated from the same unpredictable climate models. Considering environmentalism and enviro-skepticism as ideologies without science is most strongly supported by their response to geo-engineering. The people who do not trust the predictions of the models, the skeptics support geo-engineering. The activists who are so sure of the model predictions in claiming the onset of catastrophes oppose them.

In the previous paragraphs, the major impediments to strong action, the lack of motivation, the limits of economic and social action and the unreliability of mega projects have been alluded to. It is true that CFC's, (chloro floro carbons) used among others by the refrigeration industry were globally replaced. This was mandated by the damage they caused to the ozone layer that screens the earth from harmful ultraviolet radiation emitted by the sun. Unfortunately, unlike the replacement of ozone depleting CFC's, carbon dioxide does not have a replacement. There is no current robust technological solution for environmental action.

However, as mentioned earlier, the lesson from the failure of the predictions made by Malthus in his historical paper are very illuminating. Malthus predicted an exponential increase in population that could only be stopped by a calamity. The real population of the world is projected to stabilize at about 9-10 billion by the year 2050. This reflects a change in life style that was not expected by Malthus

at the time of writing the article. Female literacy, female education and increased wealth are among many parameters which correlate with decreased number of children. However two issues are critical. One is the easy availability of technological solutions that are safe and easy to implement. Birth control pills, condoms etc are easily available. The second is the dissemination of knowledge.

Personal ambitions are certainly most important and factors such as education empower the individual to think beyond merely having a family. However, awareness of the un-sustainability of a continuously increasing population played a key role in inducing people to restrict their families. Recent evidence shows that people with extremely orthodox religious views have significantly larger families uncorrelated with the usual socio-economic factors.

One key aspect of extreme religious views is the unwillingness to accept sustainability limits. Even while accepting or even extolling human stewardship of the environment, religious thought refuses to accept limits on population and in particular technology as a means of limiting the population. This supports the claim that awareness had a key role in stabilization of human population. It is thus quite relevant in mitigating global problems.

We now chart a program for mitigating the risks of climate change that could possibly work like the uncoordinated human effort which has at least limited the population growth rates. The expression of environmental concern forms part of the required awareness. By a simple process of quantification one can evaluate specific personal life style choices that are available with current technology.

The advanced society with its high human development index, desired by most of those who are not part of it currently consumes at least 125kWh of energy per day per head. This is termed a minimum since some societies with the same level of human development currently consume up to twice the value. Since every country has poor and rich people, the above is an average between the guzzlers and small consumers within the society. Even for countries with low human development indices, the low average consumption figures are similar averages.

Now let us consider how this energy is spent. Each and every human activity has an energy cost. Not all activities contribute equally to the above mentioned energy required for the desirable life style. Significant contributors which dominate consumption can be easily identified. These are transportation and space conditioning. When a person travels by car he spends energy not merely for moving his own weight but also the vehicle. When he flies he adds the additional energy required to keep the airplane in the sky. When we talk of space conditioning we do not envisage only the space occupied by the human but a much larger space in the house or office. These together account for more than a third of the energy requirements and consequently have the most scope for conservation. Other activities are too minor to quantitatively matter.

Further, the energy consumed by one intercontinental flight of approximately 15000 KM in the economy class would be equal to the energy consumed in one year by commuting 50 KM per day in a car. Additionally, a person can comfortably travel 200,000 km in a year by air but only a tenth of it by car. It is possible for a person to consume energy orders of magnitude higher than the average by profligate air travel. While the rich in using personal aircraft or huge mansions may consume significantly higher energy, numbers being small, the overall contribution to the society is notional rather than real. These few seemingly random observations immediately illustrates the kind of changes in life style, consciousness of which would probably be required as more inhabitants of planet earth desire and acquire lifestyle with high human development. On the other hand, paying attention to every single consumption for example by switching off mobile chargers since "small drops of water make a mighty ocean" does not contribute much. Statements of human wisdom are fine but quantification is important as this whole monograph attests. For individuals enjoying the desirable high human development indices and these exist even the in the underdeveloped societies, the most rational decisions would consider limiting the above mentioned dominant contributions. The key contribution towards any progress in this direction will emerge from simple quantification based on dissemination of knowledge and rational personal decisions about energy saving and utilization. Emotional environmentalism will not contribute significantly.

Technology is necessary for mitigating the possible climatic consequences of increasing carbon dioxide levels. However, potential scientific and technological changes cannot be demanded by their desirability, human effort and societal support for the effort. This is a general truism about technology and science and applies to climate change too. Societal action through politics, economics or social norms are probably required but their success cannot be superior to their success in other areas of human endeavor. Even the consequences of the rational advice to limit energy use for travel and space conditioning mentioned above are economic decisions and significant decrease in consumption at the level of the society will have unpredictable consequences.

In the mean time the inescapable fact remains that the carbon dioxide concentrations in the atmosphere are continuously increasing and are now large enough to be significant in altering the complex system of interacting processes of the living beings and the earth that is called the climate. The complexity of the system ensures that humans are left with an environmentalist's dilemma, knowing that catastrophes are possible but unable to define the probabilities of their occurrence or device global mitigating mechanisms.

XII
POLITICAL ECONOMICS WEALTH AND EQUITY

XII.1 Quantification of economics

Quantification is inherent in all business transactions. The first written records available to archeologists are bills and records of commercial activity. However, ancient economic ideas regarding taxation, ownership, etc be they of Aristotle (~350BC) or the Indian scholar Chanakya (~300 BC) were not based on numbers. The post Newtonian desire to create a mathematical model for economic activity started a mathematical analysis of numbers representing economic activity and thus modern economic science. Economics today is more mathematical than even physics. However use of mathematics seems to elicit different responses from physicists and economists. While a physicist revels in the success of mathematical physics and wonders why nature is mathematical, there are strong claims of overuse of mathematics by respected economists themselves. Such criticism by the practitioners themselves is not seen in astrology or homeopathy and demonstrates the genuine scientific aspirations of economists. Support of economic science is claimed for the strongest political ideologies, the *laissez-faire,* free market ideology and Marx inspired communism. Whether economics really is a science remains disputed

despite the annual award of a "*Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel*". The free market and communist ideologies are routinely attacked with arguments of their being contrary to moral sentiments and/or human nature. However, the goal of the present discussion is to know the strength of the "economic science" behind the ideologies and such arguments are not employed here.

Economics is described as an attempt to understand optimum and rational basis for allocation of limited resources among much larger needs. Defined thus everything, spiritual aspirations, love and affection, desire for excellence in any field is economics. Time, physical and mental capabilities of the individual as well as societal resources are always limited and cannot be sufficient to realize all desires. It is easy to dream but difficult to implement! This illuminates the relationship of economics to social sciences as nothing else can. Trying to divorce economics from other social issues is actually quite futile. Human wants, even if they are physical and "commercially available" are always associated with moral sentiments. Modern economics discusses allocation of resources that can be assigned a monetary value to cater to human needs that can be satisfied by physical entities. Thus as with other sciences, there is an attempt to divorce the "is" from "ought" as a first step towards creating a science of economics. One apparent consequence is renaming political economics as "macro-economics". However this project of creating an objective science is even more difficult in economics than in biology which similarly studies living beings. Despite these limitations, in as much as a strongly mathematical economic science is used to support the ideologies, the usual question of how well we know this is discussed.

XII.2 Laws of market economics and their mathematical formulation

The basic economic laws, those that are considered fundamental by the dominant, classical/neoclassical school of economics emerge almost naturally from the definition of economics as the study of allocation of limited resources on unlimited desires. The first law

explores availability of resources. In any reasonably complex scheme, some resources will be more appropriate for producing a specific product. Using a typical example used by economists, some land is most suitable for growth of wheat. If the production has to increase, land that is marginal for the growth of wheat, perhaps more suitable for growth of maize has to be used for growing wheat and this would enhance the cost of production. Also this would significantly reduce the production of maize. This is termed as increased opportunity cost.

This is generalized into an expectation that if the price of an object is higher, production could increase since resources not perfectly suited for the product can also be employed. If the price of wheat is high, use of land marginal for production of wheat is justified. In principle, the available resources can all be categorized according to the suitability and hence as production increases, the opportunity cost required to supply the production will continuously increase. Correspondingly the variation of supply with cost will be a curve and is referred to as the law of supply.

The demand or desire for any specific object intuitively depends on the price. It is reasonable to expect that a customer with limited resources would reduce the consumption if prices increase. A more detailed description of this demand curve is provided by the concept of marginal utility. Just as many resources can be employed for the supply of a specific product, most products serve more than one function for the consumer. The various functions can be categorized by the consumer as more or less important. The use of water for drinking is much more important than its use for watering a rose bush. When a large quantity is used, all the more important functions are satisfied first and the last unit is employed for less desired outcomes. The "marginal utility" as it is referred to is low. This is generalized into the law that the marginal utility decreases as the quantity used increases. This in turn leads to the law of demand where the demand decreases as the price increases. When the cost is high only the important uses are fulfilled. If water was costly we would not waste it by watering a rose bush but reserve it only for drinking. Once again the curve is the result of categorizing the various uses in decreasing order of importance, something that is possible in principle. These curves are shown in the figureXII.1

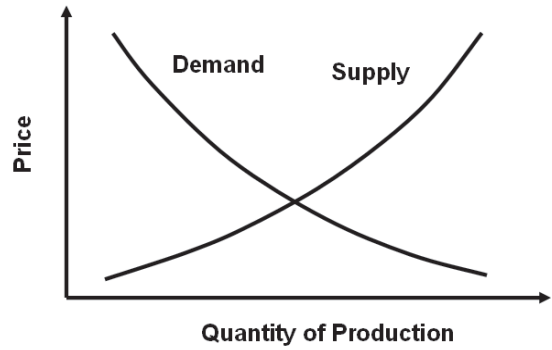

Figure XII.1 Variation of supply and demand

Once the logic of the decrease in demand and increase in supply with increase in prices is accepted, it becomes obvious that an equilibrium price would exist as long as both the buyers and producers are free to take independent decisions. No planning or collective action is necessary. In their absence, the attainment of equilibrium has been attributed to an "invisible hand". Since exchange is a voluntary action on the part of the supplier and the consumer, a deviation from the equilibrium which benefits the supplier would lead to loss for the consumer and vice versa. This logic becomes another primary principle of economics, that of "Pareto Efficiency". A system is Pareto efficient when the system cannot be altered to provide benefit to one without a cost to some one else. That a system of free trade is Pareto efficient is thus not a surprise since it is only a restatement of the voluntary exchange principle. If both individuals are "free agents" participating in a trade, they are both presumed to be satisfied with the situation and any modification of the terms of trade cannot but benefit one over the other. Thus by definition free trade is not a sum zero game, while any alteration of free trade is a sum zero modification. When evaluating the strength of the science it is necessary to clearly notice that the logical statements of economics are not independent and thus the total empirical content is small.

While one can conceive of a simple exchange as being the representative basic unit of economic activity, the overall economy is much more complex. Government, companies and individuals inside the country interact among themselves and with other countries through at least three markets namely the labor market, the financial

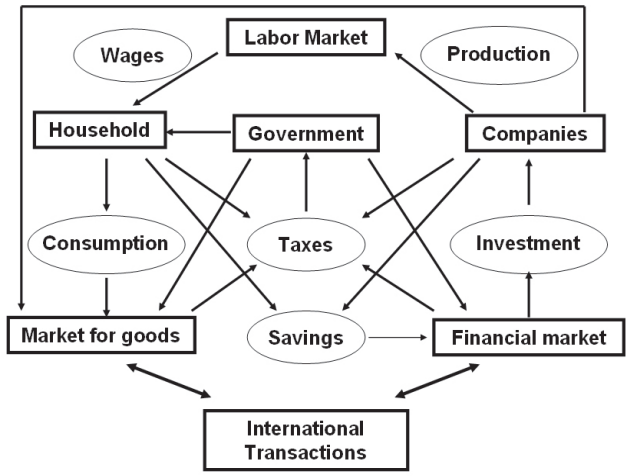

Figure XII.2 The interactions in an economy

market and the commodity market. The labor market provides a worker for a given wage. The financial market provides investment for a return and the commodity market provides materials for production for a price. Wages, prices, savings, taxes and transfers are among the various financial interactions. An example of interactions in the economy is shown in figure XII.2. Whether the equilibrium between supply and demand extends to entire economic system is a natural question for economic science.

There are no logical reasons that rule out the possibility of such an equilibrium. Economists claim to have demonstrated the existence of an equilibrium in many model systems. These models are constructed employing some idealized mathematical formalism for the interactions. Some of the sophisticated models are called "stochastic", which means that the interactions are modeled using probabilities for various actions rather than certainties. Some models incorporate features of "real rather than ideal" markets. The question of whether the resultant models are realistic enough to reflect the real economy is contentious. One group called the neoclassical school, contends that the complexities have largely been taken into account and a realistic nearly free market solution is not only desirable but is also logically consistent. Several other groups question this dominant view and claim that the complexities are more significant and that

not only is a nearly free market not desirable, it is logically flawed. That such a disagreement exits is quite well known.

The difference between these complex models of economics and those encountered in the earlier chapter on climate modeling is obvious. Unlike the climate models, physics cannot suggest functional forms. However, as in the climate models, the mathematical functions representing the various interactions have to be continuous functions which are differentiable. This essentially allows the rate at which the variable is changing to be determined and permits identification of equilibrium. To be absolutely logical, no economic variable can be a continuous function. An individual could consider the utility of two apples to be equal to three oranges but a question of equivalence of say 1.9995 apples to 3.0005 oranges is silly. Utility cannot be ascribed to smaller and smaller units indefinitely. While this is a logical objection, this in itself does not rule out the use of continuous functions as a reasonable approximation.

More importantly, at least some of the interactions in a real economy are not represented by a balance between marginal utility and opportunity cost. The ideal relations of the marginal utility and the cost of production have been described. Even the most elementary exposition of economics provides examples of deviations from the ideal. For example, when items are extremely important such as food, increased cost does not cause reduced demand. Similarly even if the cost increases, supply of paintings by a great artist does not increase. Such variations are incorporated in the models of the economy as elasticity of prices. However, empirical evidence for the laws of supply and demand is rather scarce. When attempts are made to empirically confirm these basic laws the results are certainly not conclusive. This once again results in bitter arguments based on desirability criterion mentioned above.

The laws of economics are valid when other things are held constant. That technological progress reduces the cost is common experience. Thus the law of supply, increasing supply with increased cost can be empirically observed only when there is no technological change. Since technological progress is continuous, at best empirical data supporting the law can be obtained over short intervals of time.

Similarly the key aspect of a modern economy is specialization of labor. Thus the increase in supply of labor will crucially depend on training and education. These parameters are outside the purview of the dictum "*ceteris paribus*" (keeping other things constant). To validate these laws is extremely difficult if not impossible.

At the macro economic level economic science is also limited by the non availability of proper economic variables. There are significant problems in defining even simple variables such as Gross Domestic Product (GDP) and inflation rate. For example, GDP as it is normally determined goes up if a large number of people make excessive purchases, take large home loans or overspend on their credit cards even if these are irrational economic decisions. GDP also increases when goods are produced but not sold. While even Adam Smith distinguished between productive and nonproductive labor, GDP statistics do not. GDP includes cost due to air pollution and money spent on cigarette advertisements. Obviously, spending as reflected in GDP does not constitute prosperity. Another economic variable often employed for macroeconomic modeling is savings rate. There is not even a standard methodology across nations and multinational organizations for calculating it. Inflation or increase in prices of the economy are officially measured and announced in most countries but the very process of determining something like the price index is clearly arbitrary and the numbers are vigorously and sometime acrimoniously contested.

The absence of fundamental theoretical basis for economics unlike physics is well known. Current economic discussions are so complex and mathematical that it is difficult to know if the models are sensible and usable. This is once again the same situation that we encountered in the earlier chapter on modeling of the environment. However, we found it possible to evaluate the strength of the different claims without being lost in the complexity of the models. Economic laws will always be imperfectly known. How far these imperfections place doubts on the strength of economic science and the problems of interpolation and extrapolation described in an earlier chapter remains contentious among experts supporting opposing ideologies. The most amusing aspect of the comparison is the tendency of the same individuals who doubt the utility of the climate models to accept

the economic models without any doubt though they are far weaker. The interactions in climate models are based on physics the economic laws are not.

XII.3 Contingent limitations on the mathematics of market economics

In an earlier chapter, the contingent limitations of physics on biology and medicine were discussed. The laws of physics, as has already been said are completely unrelated to the variables and laws of economics. Obviously there cannot be any contingent limitations of such a nature. It is however common to come across terms like "equilibrium" in economics and it is pertinent to see if physics can still be of some help in understanding the limitations of the use of such ideas in economics. Moreover, since extremely complicated mathematics is being employed, it is pertinent to look at economic science in the light of those discussions.

Equilibrium in physics is linked to the idea of ergodicity and Liouville's theorem. Within physics there are many ways of discussing these principles but for our present purpose, the appropriate way to understand the principle is to consider a sample of gas consisting of large number of molecules. As mentioned during the discussion of Boyle's law, these are randomly moving in various directions with various velocities. Assume that the velocities of the various particles

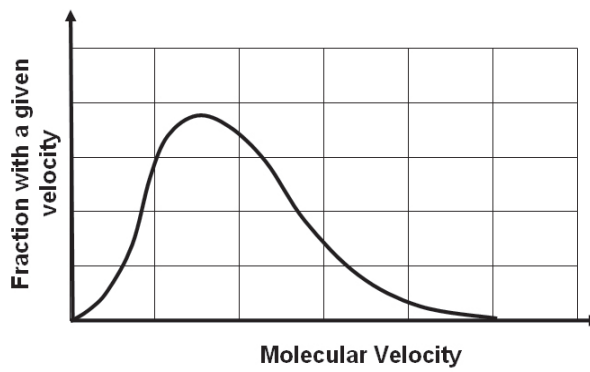

Figure XII.3 The Maxwell-Botltzman distribution of the velocities of gas molecules

at a given instance are known. Now one plots the percentage of molecules (or equivalently the probability of a molecule) having a particular velocity. This is shown in fig XII.3 and is called the Maxwell-Boltzman distribution of velocities. There are very few particles with velocities close to zero and very few with an infinitely high velocity. This leads to a curve which looks like the black body radiation but is quantitatively quite different.

The collection of gas molecules is said to be ergodic. This assumes that if one particle is monitored continuously its velocity continuously changes as it collides with other gas molecules. The duration for which it has any given velocity can be determined in principle and the probability for various velocities determined. The ergodicity principle is the statement that once again, the probability of the velocities would be a Maxwell-Boltzman distribution. The fraction of molecules with a certain velocity will be equal to fractional time for which a single particle under observation has the same velocity. One can immediately discern a fundamental equality between the time evolution of one sample and an ensemble of identical particles. It is their equivalence that provides confidence in accepting the average properties of the collection of gas as determined by observations as true and stable or in equilibrium.

Just as one talks of a gas, a large collection of molecules being in equilibrium, one talks about the economy, a collection of large number of exchanges of goods and services being in equilibrium. The description of physics immediately highlights a major limitation of equilibrium in the case of the economy. The equivalence between the time average and average over all transactions of the variation of price with demand would not carry much conviction.

A mathematical curiosity sheds some light on the complex mathematical models of the economy which are used to demonstrate equilibrium and Pareto efficiency. Any mathematical function of many variables will have many minima. To understand this, consider that the values of the various variables are changed and the corresponding value of the function determined. If there are two variables one can imagine a surface like a rubber sheet suspended across a room. Each point on the floor corresponds to one pair of values of the two

variables. The height of the rubber sheet will then be the value of the function. Now it is easy to imagine that the sheet is not a smooth surface. It may look as if stones have been placed in it at various points pulling it down. These are minima, values of the variables corresponding to which the value of the function is locally lowest. Near each of these points, any change in parameters would result in the value of the function increasing. Another way to visualize this is to imagine a small lake. The bottom of the lake can have any complex shape. There could be many small shallows. Any movement from the bottom of these local shallows in any direction will be upwards. The slope is minimum at the lowest point and the technical mathematical terminology is that the derivative of the function is zero leading to a stable situation or an equilibrium. If there are many minima, similar to there being several deep pools, the lowest is called the absolute or global minimum and the rest are local minima.

The mathematical curiosity we now consider is the function represented by the cube root of unity. That is the solution of the equation $a^3 = 1$. There are three roots or minima which are 1, -1+i, and -1-i (the square root of -1, is normally represented as "i" in mathematics). Let us consider extending the analogy of the lake or basin and imagine that the three solutions are the three minima and ask how the area around the minima would look. This is easily

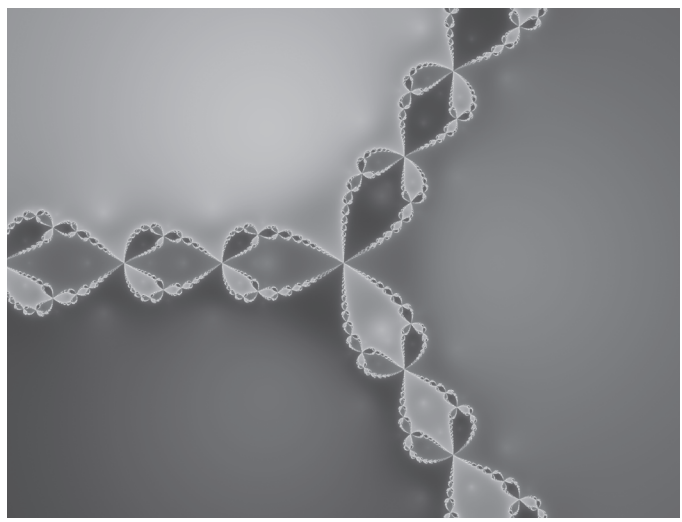

Figure XII.4 The basins of attraction of the cube roots of unity

answered by considering what path a drop of water would take when it is placed close to one of these minima. From all regions around the minima water flows to that minima. The regions for each minimum can be shown shaded in a figure. In the case of the cube roots of unity, there are three regions and a boundary must exist between these local regions. At this moment we are not sure where the water would flow from the boundary.

The boundary has been plotted by computer and shown in figure XII.4. The boundary between the three regions is very interesting. A drop of rain falling at the extreme left of the picture, instead of flowing into one of the two nearby minima actually comes to the third or far away minimum. This is indicated by the various shades in the boundary. This is a fractal or self similar structure. Expanding the small regions will show much more complexity. This is similar to the deterministic chaos that was discussed in an earlier chapter and again in relation to climate models in the last chapter.

Admittedly the description of the fractal geometry of the boundaries of basins of attraction of the cube roots of unity as this is described in mathematics has only a tenuous relationship with economic models. The reason for this digression is not to question the mathematical foundation of the various models of the economy. Not even to question the accuracy of the simplifications or even the moral desirability of Pareto efficiency. It is to highlight the complexity involved in even the simplest mathematics. While discussing the climate models, deterministic chaos was described not to claim that climate is chaotic but to show that complex systems move from one stable condition to another in a random and sudden ways. Similarly, this mathematics is a warning that in a complex system, local movement towards one of the available minima does not guarantee a reduction of distance towards that minimum. If one is on the fractal boundary, moving locally towards one minimum will lead you to another. This gives a warning that every movement towards a free market, from a current situation is not always beneficial merely because the free market system is the Pareto efficient equilibrium. The consequences of any such change from a current situation far away from the ideal market would be impossible to confidently predict even in mathematically simple systems.

XII.4 Time as a variable in economics

Given the open admission of physics and more often Newtonian physics as the basic inspiration for mathematical economics and the use of differential calculus in the enterprise, it is very surprising that time has not been quantitatively used as a variable in economic models. The most important success of Newton can be attributed to the realization that the second derivative of position with respect to time, the acceleration was the most critical parameter in physics.

Claims that models of the economy have been refined to incorporate corrections to the ideal free markets in order to ensure that they are a realistic representation of actual markets were mentioned earlier. In discussing these models, often claims such as the following are made. "While in the short run it is possible for an individual firm to make an abnormal profit, in the long run such abnormal situations caused by imperfect markets are not possible". Similarly, "While the problems associated with reduced aggregate demand for goods in the down turn of a business cycle, as identified by Keynes are valid, the economy follows the neoclassical theory in the long run". The key worry is that the time scales involved in such economic jargon are never defined in absolute terms. It is not clear if the short and long term refers to days, months years or decades. Keynes said, "in the long run we are all dead". In any case, there is no clear theoretical basis for assigning any absolute time scale leading to clear disagreement between economists. One suspects that the oft repeated joke "if we place an economic problem in front of five economists one obtains six different answers" is rooted in this inability to quantify time.

The most common statement in economics that quantifies time is the market folklore that movements of a stock or currency will look alike when a market chart is enlarged or reduced so that it fits the same time and price scale. An observer cannot distinguish whether data concerns price change from week to week, day to day or hour to hour. This statement immediately resembles the description of a fractal. A simple example of a fractal called a Sierpinski Triangle is

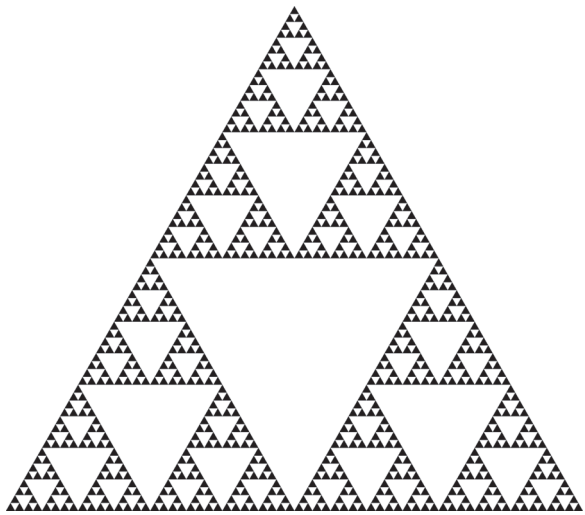

Figure XII.5 The Seirpinski Triangle

shown in figure XII.5. (It is much easier to appreciate the self similarity in this than the other examples cited earlier). The similarity between the smaller and larger parts of the figure is quite obvious. This triangle can be constructed using the smallest part very much like a children's building block.

A similar highly simplified picture of the variation of market prices has been constructed by Mandelbrot using a small segment of the up and down oscillation of a market price. This is shown in figure XII.6. The first curve has three segments, an increase followed by a decrease and then a final increase. In the second curve, each of the three individual segments is replaced by three segment lines similar to the first curve. The initial curve is also visible. The second curve has 3×3 =9 straight line segments. In the subsequent curve each of

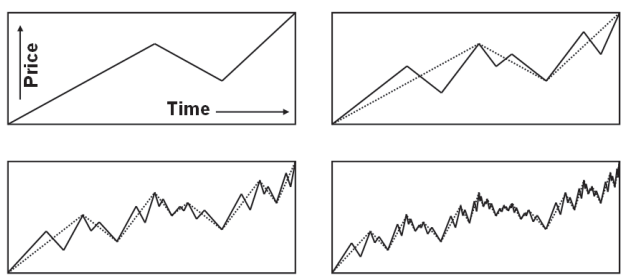

Figure XII.6 Creating a fractal curve of price variations

these is replaced once again by a three piece segment resulting in 27 segments and so on. The initial segment is scaled down to smaller and smaller length scales and added resulting in a complex curve. As one goes through this seemingly meaningless exercise, time for one of the segments in the initial up-down-up market activity is reduced representing rapid market activity. This results in large fluctuations in the market prices. More importantly for a bird's eye view description of economics that is being provided here, the resultant distribution of market prices is no longer Gaussian. This is called "fat tailed" or "enhanced tails" distribution. An event with a deviation ten times the standard deviation(s) is extremely unlikely with a Gaussian distribution. As discussed in an earlier chapter, there is a 99.7% probability that the value is within a band of 3s and the probability falls of very rapidly after that. The probability for 10(s) is actually $\sim 10^{-23}$. In these fat tail distributions, the probability of events with large deviations does not decrease but remains finite for very large deviations, as seen in figure XII.7. Such distributions are encountered in selected areas of physics as are fractal descriptions. Significantly, when large changes in the market are observed, statements are made that these were highly unlikely and therefore not expected. Events which are considered extremely improbable occur so often in the real economy that several famous economists believe that the distributions in economy are non-Gaussian. Mathematical description of such distributions is most often not possible, restricting the possibilities of incorporating them into complex models of economic activity. As in the earlier discussion of ergodicity and basins of attraction these physics descriptions provide food for thought than proofs.

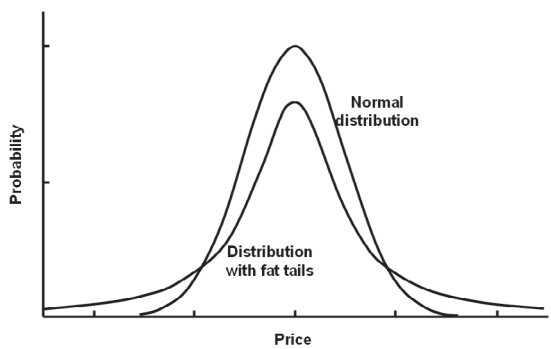

Figure XII.7 The fat tail distribution

Market activities (actually stock market or financial market activity) occur over a wide range of time scales. Fruitful economic activity, investment in real assets and production using skilled labor occurs over a much more restricted time scale. It is difficult to imagine setting up a unit for production of even the simplest object necessary for a consumer in less than a month and the most complex project in human endeavor normally lasts only about 10 years from serious planning to execution. Thus one has a time scale of approximately 5 to 500 weeks or a factor of 1 to 100. The time scales required for acquisition of skills by labor to increase its supply in accordance with demand also fall squarely within this time scale. It takes typically a year to train a machine operator and about five to train an engineer.

Barter without money is not very efficient since the double coincidence of wants, the buyer desiring a specific product physically meeting a seller of that same product and willing to negotiate the price is logically rare. In a complex economy, financial and stock markets are similar reasonable requirement to efficiently link money for investment and entrepreneurs. Similarly commodity exchanges are a logical requirement for providing estimates of future prices of raw materials and indirectly insurance coverage to manufacturers.

A question naturally arises if the fluctuations in the so called "market activity" represents anything real when these take place at time scales much different from those of the productive economy. Clearly fluctuations in the really short time frame, intra day for example cannot influence the productive economy in any way. This is accepted and the fluctuations are seen primarily as a barometer to gauge response to government actions.

It has been empirically observed that the really long term (25-40 years) return in the stock market is approximately equal to the GDP growth that has been independently determined. This is understandable since with all its limitations, some of which were discussed, GDP does represent the real goods and services available. Clearly operations at short time scales in these markets are unproductive activity as opposed to the basic economic activity which continuously enhances services and goods. Large number of transactions and easily available data make these "markets" available

for economic analysis. However, is the activity really representative of economic activity? Can mathematical analysis based on this data strengthen economic science? Moreover, the extremely short term fluctuations do not average out as had been discussed with the help of the fractal model. When very rapid transactions are permitted, fluctuations in various time scales appear and these overlap the time scales of productive activity.

Fluctuations in the time scales of productive activity, typically of duration of one year or so can adversely impact operation of industries leading to their failure. For example, in recent years, price of oil fluctuated from 50$ to 150$ and back to 50$ over one year. Such a change is large enough to cause failure of firms. But the fluctuation may be a consequence of permitting market activity at very small time scales. Are firms that fail logically "dead wood", inefficient concerns as assumed by the free market economic theory?

Another important aspect is the magnitude of this speculative component in economy. The influence of the magnitude of speculation has not been included in the models of economic activity. It is entirely possible that specialization and division of labor may result in a local economy, be it of a city or even a country being dominated by these "sum zero" activities. It is not clear if the much touted models of economy take these factors into consideration.

So far the logical and mathematical foundations of the free market economic model were investigated to determine the strength of the science behind free market ideology. Only a specific question, whether there is a reasonable scientific support of the ideological position that every movement towards free markets is always justified was examined. "The road to serfdom", is a common fear voiced by economists about the "eventual" consequence of moving locally away from the "free market ideal". This argument is used to oppose all movement away from the free market principles. The discussion in the previous few pages showed that while the basic laws of economics are reasonably logical, extrapolation to the complex economy is not quite justified. These weaknesses in economic science are obvious without reference to the deeper questions of ethics and desirability on which most social issues are to be eventually discussed.

XII.5 Economic models and microeconomics

The arguments above question the scientific basis for the ideology which supports every local movement in the direction of a free market as good for the whole economy and opposes every step away as a slippery road to disaster. There is another part of economics called microeconomics which attempts to provide solutions for smaller and simpler problems. Use of this condescending term is the converse of respect given to terms like "pure mathematics" and "fundamental physics".

Microeconomics may provide advice to pharmaceutical companies of the risk of malpractice suits outweighing potential profits from the sale of a drug. It may suggest a counter intuitive strategy for pricing tickets based on profit maximization graphs or statistical forecasting of the price of fuel to enable an airline to take insurance against price fluctuations. Our argument does not extend to such use of economics. Logical laws of economics are more valid in the smaller less complex area. They have been deduced from logical analysis of small transactions. It may be possible to model the empirical data more effectively for small deviations from the ideal behavior. Most importantly, these statistical mathematical models are not extrapolated to other areas. Their use is justified by the experience of a number of potential users and may be dropped if found not attractive. Even if the mathematical decisions are not justified, an equivalent of the placebo effect could operate in economic sphere too. Just like the psychological boost offered by an ineffective medicine, the support of "knowledgeable" economic advisors could increase confidence which is very important for economic success. This alone could lockin the use of such advice since "everyone is doing it". This dichotomy between the micro and macro economic situations is a very important pointer to the essential difference between applied disciplines like engineering and medicine on one hand and applied economic science on the other. In engineering and medicine, as has been discussed earlier, fundamental theoretical physics acts as a limitation. While physics cannot proactively guide these disciplines, they cannot operate counter to the principles of physics. The laws of economy have no such relation with fundamental physics nor do most social sciences.

XII.6 Alternative economic ideologies

The economic analysis of Marx stands as a complete antithesis to the free market and is at the foundation of the other most important ideology. This holds private property as the core of all evils. Free market economy is perceived to have an inherent internal logical contradiction that will inevitably lead to its complete failure. While not every socialist is a Marxist, the core slogan of the socialists, "from each according to his ability, to each according to his need" does resonate with the human emotions quite as well as the logic of trade and voluntary exchange. While questioning the utility or desirability of a government sponsored transfer of resources to the needy, a champion of free markets would suggest "individual charity" as a more "efficient" mechanism for satisfying this requirement. It is interesting that there is no attempt to explain why "double coincidence of wants" would be much more efficient for charity than barter. After all it is illogical to assume that the despondent individual requiring charity would immediately and easily find a giver with the required resource and immediate charitable disposition.

The current approach would simply look at one easily understandable feature of Marx's analysis. Just as most of the complexity and detail of the free market has been ignored we decline to look at the entire Marxist ideology in all its detail. We discuss one particular version of the analysis which is simple enough to be discussed. It appears to provide the essence of the Marxist argument. In line with this description, an abstract idealized system of capitalism is constructed. It can then be logically demonstrated that it is headed for disaster. The claim that any real system of capitalism which is not even ideal will collapse even faster follows.

Marx essentially argues that increase in demand for labor would increase wages and decrease profits creating a demand for mechanization and innovation to reduce labor costs. Innovation will decrease demand for labor and naturally decrease wages and thus increase profits. This is so much a free market axiom that to call it the basis for Marxist analysis looks ridiculous. Given the animosity between Marxist and free market thought, it is really surprising how

close Marx's assumptions are to absolutely free markets. However Marx assumes, or more correctly extrapolates this logic. Due to the consistent drive for innovation and replacement of labor, wages decrease to the level of subsistence. Clearly, in an ideal market, there is no logical limit to the decrease in wages or improved innovation other than the minimum subsistence wages for labor. Below this level the laborer could not even survive.

The second assumption in Marx's analysis is strong competition between firms for increased profits that can be obtained by increased innovation and lowered labor costs. As in the free market analysis, firms that are unable to compete are inefficient and will be replaced or absorbed by more efficient units. The only additional assumption is that bigger firms are stronger. This assumption is at least not illogical.

The logical extrapolation leads to large industrial agglomerates that have absorbed smaller inefficient firms, have the resources for innovation and operate with the labor working at subsistence level wages. The absorption of the smaller units, particularly those of artisans and cottage industries, leads to labor becoming a commodity with the individual worker having to "sell" his labor in the market. This logic leads to Marx's version of capitalist society with a minuscule rich capitalist class and the vast majority being workers at subsistence level waiting to participate in the revolution. This is the inevitable destination of even the most idealized model of a free market economy and the direction of history.

Even if one does not examine the details of the rest of Marxist philosophy one can appreciate the Marxist's confidence in the logic which, it must be admitted, is as logical as the invisible hand that leads to Pareto equilibrium. One interesting observation is that time does not enter as a variable into this idealized Marxist description of the market either.

So how fast this process of agglomeration will proceed and how soon the revolution will become inevitable is not determined. Just as a free market fundamentalist applauds in principle every local movement towards a "freer" market, every Marxist sees "exploitation"

in every decrease in wages and the necessity for an overthrow of the system if necessary by violence.

In addition to logical strength, the analysis seems to be empirically supported. Since the analysis 150 years ago, modern societies have largely eliminated the artisan class and the overwhelming majority is in the labor market, thus labor has indeed become a commodity. The business cycles closely resemble Marx's description of failure of non-competitive entities. Even Marx's insistence that the true value of a product is simply the labor required for its production seems to be empirically true. If one looks at the price of any object that is commonly available, there is a labor component in production, in sale, in tax collection and spending of revenue by the government. There is the intellectual property which once again is a labor cost. Any equipment used for its production is itself an object with a labor component. Thus logically one gets an infinite number of labor components. The only non labor component is royalty or rent paid for the natural resources. In any modern complex economy this is such a small component that it can ignored. A statement that there is but one property, the intellectual property and labor would not be normally considered as a Marxist statement but seems to be empirically correct. Marxist "absolutely free market" of capitalists who have no role in production other than providing capital appears to be true for the financial markets.

But instead of the descriptive approach followed above, one can examine the strength of the extrapolation by quantitative empirical observations. Please note the distinction of this approach from the usual philosophical or descriptive arguments. As usual for the purpose of our current effort logical does not include "logical arguments in words". Consider the idea that a bigger firm is necessarily more efficient and hence will absorb the smaller firms. Is the US economy dominated by a few major firms? According to one estimate firms with more than 500 employees contribute only 24% of US GDP. Certainly they are an important part but certainly they do not have the dominance that they are accused of. Even more interesting is the actual life of a corporation or industry. They seem to grow and to die. The smaller firms, the startups as they are called seem to grow much faster and push the lumbering giants aside. No company has dominated the

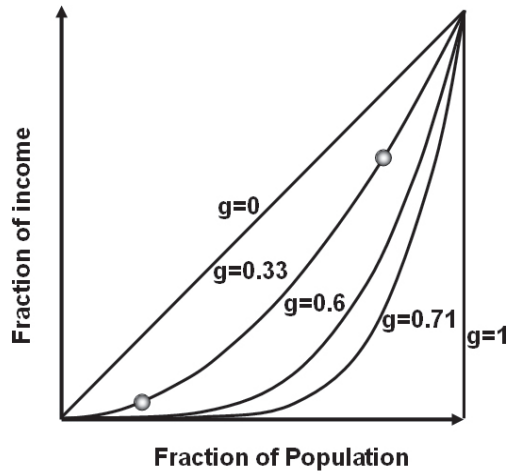

Figure XII.8 The gini coefficient defines distribution of wealth in the society

economy for even two generations. That the large firms do not move fast enough is an informal "truism" that recognizes the importance of including time as a variable.

Next consider Marx's claim that ultimately the free market will result in a situation that a vast majority will live in subsistence wages alongside a minuscule capitalist exploiters. This is mathematically described by a gini coefficient. The plots in figure XII.7 plot the population fraction against fraction of wealth or income. If there was ideal equal distribution of wealth one should get a straight line with a Gini coefficient 0. If the wealth is concentrated totally with just one individual with all the rest having nothing, one has a Gini coefficient of 1. This will be a straight horizontal line ending in a straight vertical line shown in figure.

Marx envisaged a revolution as necessary to move this to the egalitarian situation. The problem is that most western capitalist countries have smaller Gini coefficients and are more egalitarian than the underdeveloped and agrarian feudal societies. In an advanced society, a large fraction of the society is involved in work that requires specialized skills and people would not acquire those skills or use the skills unless there is a financial incentive in the form of large enough

wages. Failure of the extrapolation from logical starting observations is common to both the free market and Marxist ideologies.

It is amusing to apply the idea of gini coefficients to workers in different fields. There will be a variation in earnings among skilled professionals such as doctors but still the range would be much larger among the market managers and other "economic decision makers" where almost unlimited financial gains are possible in the sum–zero game played on stock markets. So the gini coefficient for doctors is more egalitarian than for the financial experts.

The problem of the non-productive speculative markets dominating the country's economy was already mentioned. There have been recent reports of increasing gap between the rich and the poor in the United States. Can this increase in the gini coefficient be attributed to a significant section of the population being in this segment of the economy? The economist Nassim Taleb labeled jobs as being in either mediocristan where the returns are limited or extremistan where they are unlimited. These correspond to the above labels of productive and unproductive economies. It is reasonable to associate extremistan with higher gini coefficients.

XII.7 Limitations of extrapolating economic science to ideologies

Mathematics is ordinarily considered as producing precise and dependable results. However, more elaborate and abstruse mathematics leads to more uncertain and speculative conclusions. This probably has to be added as a codicil to the great statement of Lord Kelvin about quantification. One sometimes wonders if astrologers conned the public for a long while (and established antiquity as one of their selling points) because their mathematical skills were far superior to that of the common people. Ultimately the above discussion exposes limits of illogical extrapolation in support of the favored ideology. This analysis of the logic and mathematical strength of the arguments underlying economics questions calling economics a science. Many respected economists themselves have made similar statements but the usual approach is to catalogue the various complexities in the economic system. This has resulted in

some novel approaches to economics including a new branch of evolutionary economics. Evolutionary or Darwinian explanations have become quite common across many areas of human endeavor and these will be discussed in the next chapter. In addition, "economic" investigation of many social sciences has become very popular. These will be discussed along with other aspects of social science in chapter XIV. Once the limitations of ideology are clearly delineated, the problem of macroscopic economic decision making in a society becomes an effort at reconciling mutually exclusive views and estimating the relevance of statistical evidence. This places economics correctly in the company of other social sciences, subject to all their limitations as discussed in chapter XIV.

XIII
EVOLUTIONARY EXPLANATIONS SCIENCE AND UTILITY

XIII.1 Theory of evolution

Evolution in the present description is synonymous with Darwinian evolution. In simple language this theory proposes that in an environment of limited resources, random changes in a living organism will lead to the survival of those most adapted to the environment. We currently explain this in the language of genes developed by Mendel and molecular biology which were unknown to Darwin and his writings are not always in accordance with the current understanding. However unlike the co-discoverer of this process, Wallace, he had not lost confidence in the scientific validity of the idea leading to evolution and Darwinian evolution being considered the same.

Use of Darwinian or evolutionary science as an ideology is notorious history. The summing statement "survival of the fittest" has been used to support various eugenic and racist policies in addition to *laissez-faire* economics. Curiously, Malthusian economic ideas inspired Darwin. Ironically, many of those approving the economic extension of "survival of the fittest" to economics, question the validity

of Darwinian evolution in biology and support "creationist science" an ideological position based on opposition to evolutionary science. While the early extensions of evolutionary science to realms beyond biology, such as Galton's ideas of phrenology and hereditary talent were easily disposed of as "bad science", the ideological superstructure has not completely vanished. In recent years Darwinian evolutionary science has been extended to human physiology, psychology, social customs, intellectual activity and finally economics. The strength of this science is an important issue in view of the ideological debates between "nature and nurture". While some of these issues will carry over into the next chapter, an evaluation of the strength of evolutionary science becomes an important issue discussed here, in the continuing endeavor to answer how well do we know it?

XIII.2 Darwinian evolution in biology

Darwin's "Origin of Species" was a lengthy volume of 500 pages. The vast volume of scientific research since its publication 150 years ago has not prevented strong reservations in the general public about its validity though there is no dispute among the scientists and researchers. However, as shall be discussed below it is fairly easy to realize the enormous strength of the science.

The basic scheme of Darwinian evolution is a very simple extension of the Malthusian observation that population increase would outstrip possible increase in food production. In hind sight it is absurdly simple to see this. Many plants produce millions of seeds. The number of trees in the forest is limited. Obviously only a very minuscule fraction of the seeds have actually germinated and grown. Darwinian evolution is contingent on the succeeding generations not being identical. Among the newer generation, those relatively more capable of adapting to the environment leave more successors and eventually emerge as survivors. The simplest example of the operation of this principle is the peppered moth in England. Normally this has patterns of white and grey making it difficult for predators to distinguish it from the background of lichen covered bark of trees. Dark peppered moths were a rarity, produced by random changes in the reproduction process but less likely to survive since they were more easily visible to predators. During the industrial revolution

pollution and deposition of black soot, reversed the situation with the dark peppered moth being less visible and in about sixty years the dark peppered moth became the dominant, most visible variety. The situation reversed with control of pollution in recent years and the dark moth has once again become rare. A similar example of everyday experience is the occurrence of antibiotic resistance. Not all varieties of bacteria causing a specific disease for example tuberculosis are equally affected by antibiotics. During testing, the antibiotic is approved for use since it affects the most common strain. However, when antibiotics are used by the patient, the drug resistant strains are more likely to survive. If these are not eliminated by the human immune system, the disease recurs and the antibiotics are now ineffective. This is drug resistance. The drug resistant bacteria which were originally very few are now the most numerous.

The human immune system deals with disease causing organisms in a process similar to Darwinian evolution. Killing every last virus or bacteria using medicines (essentially chemicals poisonous to the microorganism) is not possible. In such large doses, they are also poisons for the human body. On the other hand a single bacterium or virus will increase to unimaginable numbers with time. The body has special cells which identify the proteins on the surface of a germ as different from the own cells and attach antibodies. Cells called macrophages then capture these marked germs and destroy them. Obviously the shape of the special killer cell has to match the shape

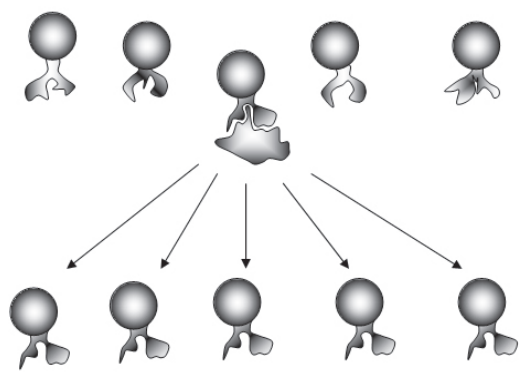

Figure XIII.1 Random mutations of the killer cells help them to match the germs

of the proteins on the germ. But the body cannot have killer cells with shapes corresponding to all possible varieties of all possible germs in the world. The killer cells in the human body undergo rapid mutation or small changes in the shape of the target biding region. They release more antibodies when they match the germs better. This triggers creation of even more killer cells with that shape. This is schematically shown in figure XIII.1. Thus, though changes in shape are random, killer cells with a perfect match to the germs are produced automatically. This process is similar to Darwinian evolution that fits the living being to the environment.

Another biological process in the human body also resembles evolution. Memory is stored in the brain in the form of a neural network. The junction between individual brain cells called neurons occurs at points called synapses. A large number of neurons connected though these junctions form the neural network. These synoptic junctions do not form at the time of storing the memory. Periodically, particularly at young age, a huge number of synoptic connections are made at random. Subsequently, those employed for memory are strengthened while the others that are unused simply destroyed. The similarity is weaker than in the case of the killer cell antibodies since there is no evidence that the synoptic connections used for the memory have been selected by their suitability. It is interesting that fairly complex thoughts are the result of pruning unwanted branches of a complex network.

Notwithstanding such observations, the claim that all living organisms on earth have evolved from a single organism by a process of random variation filtered through the environmental restrictions appears to be an example of extreme extrapolation from simple examples. Darwin's monograph became a large book when a variety of observations were included to address this essential difficulty. No serious biologist doubts natural selection and a famous biologist pressed to offer an observation that could disprove natural selection said "fossilized rabbits in the Precambrian" (earlier than 500 million years before the present). Needless to say none were ever observed. However, it becomes difficult for a lay person to judge if the small number of missing links between the apes and humans are sufficient to doubt the theory or if the evidence claimed for the links is valid.

The problem essentially is evaluation of strength of qualitative evidence rather than logical proof.

XIII.3 Fundamental physics and biological evolution

The real strength of biological evolution is physics. The age of the earth and the universe are experimentally determined quantities. The age of rocks on earth is determined by radio active isotope concentrations. In addition to the strong experimental evidence, the law of radioactive decay is not an empirical relationship. It is an integral part of the fundamental theory of physics. As was discussed before, the fundamental mathematical structure is itself based on basic symmetries and thus the measured values of the age of the earth cannot be doubted. This large time period is one of the requirements of Darwinian evolution by natural selection. The more important support emerges from the molecular structure of biological molecules. The structure of DNA has two complementary strands containing the genetic information, bound together by weak hydrogen bonds as was described in chapter IX (figure IX.1). The strong contingency imposed by physics on the various aspects of this structure, the complementary strands and the nature of bonds was described there. The genetic code

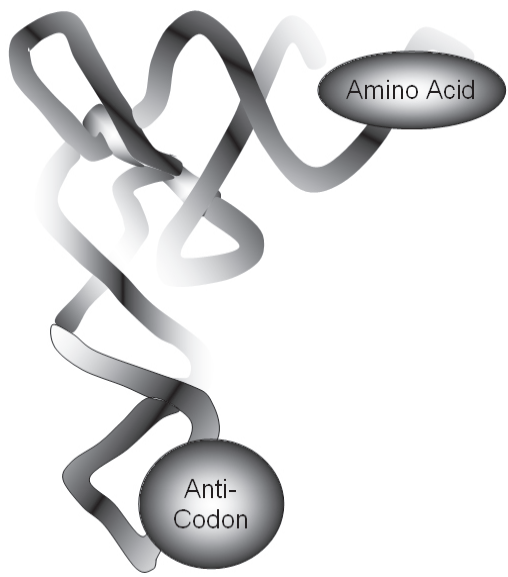

Figure XIII.2 Structure of t-RNA molecule

links the information in the DNA with specific amino acids that constitute proteins. Three nucleotides in the DNA correspond to one amino acid. For example the sequences Thymine, Guanine, Thymine (TGT) and Thymine, Guanine, Cytosine TGC in the DNA will result in the amino acid, Cysteine (Cys). There are multiple choices since choosing three out of the four available nucleotides gives sixty four possibilities while there are only twenty one amino acids that makeup all proteins. The t-RNA molecule shown schematically in figure XIII.2 plays a crucial role in the translation from the triplet code in the DNA to the amino acid. Many details such as the intermediate step involving the RNA have been ignored but are irrelevant for the present discussion. The genetic code, the relationship between the nucleotides and the amino acid is non-specific. This means that the shape of the nucleotides does not define the amino acid. The nucleotides and the amino acids attach to two ends of the t-RNA molecule as shown in figure XIII.2. This experimental fact is most important for the present. It is also experimentally confirmed that the genetic code is universal and all living organisms share the same essential code. Since the code is not defined by the shape of molecules this observation can make no sense whatsoever unless all organisms have a common origin. The process of formation of the proteins from the genetic information has been experimentally confirmed at the molecular level. This proves that the modification of the proteins cannot influence the genetic information. At the same time, the process of copying the information in the DNA during cell division does lead to small errors. One has random variations that are independent of the utility of the changes. The molecular studies confirm every aspect of the Darwinian evolution. Given the knowledge of the formation of proteins at the molecular level and every living organisms is simply a collection of proteins, a Darwinian explanation is the only possible mechanism.

Biologists employ the Linnean classification which relates all living organisms. Apes and man are very closely related, both belonging to a single "family". A snail and a man are much more distant relatives. They belong to one "kingdom" which is biologically a higher level of ordering. The strength of this classification is experimentally confirmed by looking at the genetic information in the molecules of the various organisms. The genetic information in man and ape are far more similar than those of man and snail. It is

possible to formulate the entire science of Darwinian evolution by natural selection starting from purely molecular studies without the much more difficult and qualitative description of biology. Thus absence of fossils or missing links or questions of the age of various sediments are all irrelevant to the essential scientific validity of evolutionary science as applied to living organisms.

XIII.4 Evolutionary explanations : Strength of specific examples

The relationship between fundamental physics and specific material properties was discussed earlier with the MOSFET as a representative example. It was pointed out that oxide growth rates cannot be calculated from first principles. Examples like the peppered moth described above are used to illustrate and explain Darwinian evolution. A key question in the present context is how to evaluate the strength of an individual explanation. In the case of the peppered moth, the question is how well we know that the camouflage argument is correct. Since evolutionary explanations are extensively employed outside even biological evolution, evaluating the strength of the explanations in individual cases is extremely important. Several limitations are known and discussed in research. These problems are highlighted by the religious groups to discredit Darwinian evolution entirely. In the previous section, we emphasized that any non-Darwinian explanation will not be compatible with bio-molecular process and hence this effort is futile. But it is important to understand

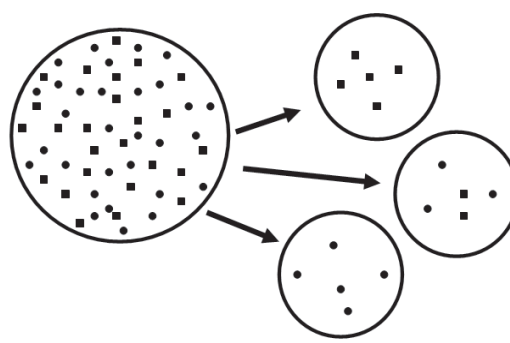

Figure XIII.3 The founder effect

the strength of specific evolutionary explanations and their utility in the human context.

The core problem in applying evolutionary explanations to specific cases is the possibility that historical accidents cannot be logically excluded. Historical accidents in the context of biological evolution can take several forms. The simplest possibility is that the fittest individual my not have survived long enough to reproduce. Thus there is a distinct introduction of chance as a parameter. The more significant contribution of chance to evolutionary changes is founder effect. When the population decreases, the survivors may not represent the population. This is illustrated in figure XIII.3. Though two distinct traits, represented by two distinct shapes are present almost equally in the parent population, chance represented by sampling could result in predominance of one trait over the other in succeeding generations. Sampling in this context is the survival of a small number or migration of a small number. This is also called a bottleneck. Thus the dominance of one of the traits in the population after the bottleneck may not result from the fitness of the successful trait but purely from chance. This can be a significant factor when for example birds spread to islands. The successful species in an island could be the progeny of the founders, a small group which has actually reached the island even if they are not the fittest. Such bottle necks in population are extremely common. For example, there is evidence to suggest that the human population has gone through a population bottleneck prior to the distribution of modern humans in all the continents nearly a hundred thousand years ago. Another possibility is genetic drift. If the population is small, there is significant inbreeding leading to the emergence of populations with inferior rather than superior fitness.

The living organism is an integrated system. It is sometimes extremely foolhardy to select one particular trait or part and discuss its evolution due to natural selection. Some of the features of an organism are extremely basic and shared by many levels of the biological organisms not merely the specific species. It is difficult to think of humans without two hands and two legs primarily because these are features shared by the humans with all primates. Having four limbs is common in almost all mammals. Discussing the same

organ, the hand further, one can investigate the evolution the fingers. Genetically hands and fingers are closely linked. Evolutionary explanations for the hand and fingers may not reflect this close linkage. A curious example is the Panda's thumb. This peculiar enhancement of the wrist bone of the panda can be attributed to the advantage of such a modification in manipulating bamboo. The same genes cause the enlargement of the corresponding bone in the foot for which there seems to be no real selective advantage. In addition to these complexities that are the consequence of the interdependence of the various parts of the organism, time can also contribute. A flying bird appears to have selective advantage over flightless birds. They have a superior ability to escape predators. But feathers of birds evolved first for regulation of body temperature and were later adopted for flight. Thus an adaptation can serve purposes other than that for which it evolved. Similarly it is possible that under changed environment, an adaptation may be present though it has no actual adaptive advantage.

In the previous chapter, while discussing economic logic one found that the strength of the micro level examples did not strengthen their macro level application. The major limitation is the absence of a linkage to fundamental theoretical structures. In the case of evolution, as discussed above, the strength of the specific explanation at the individual or micro level cannot be logically affirmed. Possibilities other than adaptation cannot be ruled out. The strength of the macro-level evolution by natural selection is much more robust primarily because of the molecular understanding and physics. Before we attempt to evaluate extension of evolutionary science to newer areas particularly to humans, we will examine some aspects of animal behavior in the light of the above limitations.

XIII.5 Extended phenotypes and animal culture

Phenotype is the name given to the observable features of a living organism. These are features that are reproduced in the succeeding generations. An old experimental study that establishes the nature of the phenotype studied the consequence of surgically removing the tails of rats. As can be expected, even after many generations, the new born rats born to these "tailless" rats had the

usual tails. The reproduction is not exact and there is a variation in the length or other features of the tail. If some of these changes provide an advantage in the number of progeny, rats with the superior features would be most abundant in the natural world. For many organisms, there are other features that are exclusively associated with the specific species, though not as part of the phenotype. Examples include, nests of tailor birds, dams built by beavers and honeycombs built by bees. The nests or dams may not be identical to one another but there is enough similarity for unambiguous identification with the species that built them. Tailor birds, beavers and bees seem to have an instinctive ability to construct these structures without any instruction. These are examples of an "extended phenotype".

Study of animal behavior also includes a variety of behavior patterns. The freshly hatched duckling follows the first object that it finds moving very much like it follows the mother duck. Wolves hunt in packs. But the two behaviors are different. Unlike a duckling, relocation of individual wolves into the wild does not result in the formation of packs. Recent studies have identified other behavior patterns among animals that have been labeled "animal culture". These patterns differ among various groups of the same species. For example, startlingly novel behavior, paternal care was the social norm in one group of primates but not in others. The behavior may also include the use of artifacts or tools. The use of twigs for fishing termites by chimpanzees was the fist evidence of tool use among non human species. Clearly the chimpanzee "knows" some amount of science to select the proper twig based on its size and absence of thorns or oozing liquids. So there is perhaps animal science in addition to animal culture. Other examples of artifacts used by some groups of chimpanzees include medicinal bandages and stone crackers to crack nuts. Washing of tubers to remove mud before eating was initiated and spread by imitation among a population of monkeys while they were under continuous observation of scientists. Differences in the vocal articulation of groups of birds, whales, dolphins and porpoises have also been observed.

Agriculture has been observed in three insect orders. Ants, termites and ambrosia beetles take several steps that closely match agriculture. While these insect species cultivate fungi rather than

plants, they select "seeds" from existing crop and transfer them to a newly prepared area. They cultivate the "crops", by regulating the moisture, temperature and humidity for maximizing the crop. They harvest crop and in many instances they are exclusively dependent on this crop for survival. More elaborate procedures closely resembling the human practices of weeding and disease suppression have also been observed.

The key question in this sketchy description is how much of all this can be considered as an "extended phenotype". Even relatively simple behavior patterns of wild animals do not constitute an "extended phenotype" though there is clearly some genetic component. Some behavior patterns are common for species spread over a vast area but are collective behaviors of packs or prides. These will not be observed in artificially raised animals. Other behaviors show significant variants that are observed only in small localized groups. Clearly, there are significant aspects of animal behavior that are not entirely determined by genes. Adaptationist logic is often invoked for many of these behaviors but clearly the distinct feature of Darwinian selection, random variation filtered through environmental fitness is at best only approximately true. This discussion of the "extended phenotype" and its limitation is relevant for the next discussion on recent advances in evolutionary science.

XIII.6 Recent evolutionary approaches

As was briefly mentioned there had been many attempts to employ Darwinian logic in examining human capabilities. Phrenology, the minute examination of the shape of the skull is one early example. Many societal decisions were made or justified based on the assumption that "survival of the fittest" is an axiomatic truth of Darwinian evolution. This has been largely discredited. However a significant part of animal behavior could be explained in Darwinian terms. The earlier descriptive ethological studies of animal behavior were made more robust and integrated into the Darwinian paradigm by the understanding of altruism, reciprocal altruism, kin selection, inclusive fitness and game theory. These conceptual advances permitted a more detailed understanding of why there are sterile castes in insects, how animals fight for disputed territory and why animals

give warning calls about predators when the risk for the individual giving the call clearly increases. This understanding of animal behavior resulted in a renewed effort to understand human behavior from an evolutionary perspective.

Just as Newtonian physics had an impact on the economic thinkers, both Lamarckian and Darwinian evolution were significant stimuli for early thinkers in psychology and other social sciences. This tendency has continued and recent evolutionary studies of humans are classified loosely as sociobiology, human behavioral ecology, evolutionary psychology, memetics and gene culture co-evolution. Thus one observes a gradual inclusion of human physiology, psychological traits, social behavior and finally intellectual knowledge including economics as subjects of evolutionary scientific analysis. Some of the explanations are examples of very good science.

Arguably the most interesting physiological trait of humans that has an evolutionary explanation is the emergence of lactose tolerance in adults. Normally adult humans as with other adult mammals do not have the capacity to digest milk. This ensures that the elder child does not compete for mother's milk with the younger who has not been weaned from mother's milk. But adult humans in pastoral societies can digest milk. This interesting variation in genetic makeup is attributed to the added survival advantage for growing children with this capability due to the easy availability of milk from cows or goats.

Clear evidence of increased frequency of occurrence of the modified genes responsible for adult lactose tolerance has been observed in pastoral communities confirming the adaptationist explanation at the molecular level. Another example is the observation of increased frequencies of the genes for sickle cell anemia among African tribes with a history of yam cultivation. Clearing of the forests to grow yams resulted in increased threat of malaria due to stagnant pools of shallow water in which mosquitoes breed. People with two copies of this modified gene have sickle cell anemia, a dangerous ailment but people with just one copy are not usually affected and are resistant to malaria. In the environment of malarial attack, the loss of fitness due to the disease is counter balanced by protection against

malaria and higher frequencies of the genetic change are observed in those communities.

The recent experimental ability of decoding the entire human genome has enabled identification of geographic or ethnic variations of gene frequencies. For example, increased type 2 diabetes susceptibility was observed in Polynesians. The founder population of Polynesians migrated from Asia. Increased type 2 diabetes susceptibility is also associated with capability of sustaining cold stress. This would have been an adaptational advantage to the founders who made the long journey in boats. In such cases the evolutionary explanation is not controversial. However despite the name of "gene-culture" co-evolution, most of the established examples refer to earlier stage of human development and small founder populations.

As one moves from pure physiological features and gene distributions, to psychological and behavioral issues, the current evolutionary scientific analysis covers many diverse issues. The almost universal and extremely strong fear of snakes is often attributed to an advantage of such fear for survival in early humans. Fear of snakes extends to at least some primates. Investigation of the human ability to guess the probability of events has been investigated extensively from an evolutionary perspective. The results show that when the events are described in words, humans make severe errors of judgment similar to the more familiar optical illusions. For example, a person was described as being "single, outspoken and very bright. A philosophy student deeply concerned with social justice and participant in antinuclear demonstrations". When asked if the person was likely to be (a) a bank teller or (b) a bank teller and active in the feminist movement. 86% answered (b) though logically the second choice is a fraction of the first. Some limits on the validity of traditional knowledge were discussed earlier. Such evolutionary research on the limitations of human ability to estimate probability of events further questions our confidence in accumulation of useful traditional knowledge through collective experience.

An interesting example, more relevant to our present analysis of the strength of evolutionary explanations themselves is the study of the ability of humans to detect violation of conditional rules. A

simple conditional rule is the legal age for drinking. It has been observed that the ability to detect violation of these rules is extremely high when questions are phrased in the social context. When the same problems are worded in abstract terms, the ability to detect is drastically decreased. This is attributed to the evolution of a capability in the brain for detecting cheating, a capability that is very relevant in the evolution of humans. Evolved psychological mechanisms in humans, in domains such as language, mate choice, sexual behavior, parenting and so on are postulated and investigated. Human psychological and social behavior ranging from polyandry, bride wealth, mate choice or homicide is investigated either as a psychological module or as behavior that evolved due to the environmental constraints.

The attractiveness of females as measured by the waist to bust ratio was investigated. Similarly efforts are made to investigate the role of fluctuating asymmetry. This measures the difference from symmetry of several biological traits such as arm girth (checking if both arms have the same girth). The observed changes have to be ignored if they correlate with other causes such as hard physical labor with one arm. High asymmetry is related to low survival and hence mate choice could be expected to reveal this "desire" for symmetry. This is typically investigated by statistically examining the number of sexual partners (in western societies) or the preference of females for the "scent" of men with low asymmetry.

As a final frontier of evolutionary explanations, the intellectual and cultural output of humans is classified as "memes" named to invoke the relation to genes. Since the phenotype including the extended phenotype is claimed to be the creation of the genes to ensure their own survival, the survival of various cultural traits is then described as the Darwinian evolution of the memes. In different research efforts, everything from religion to science is described as memes.

The quest for quantitative and testable conclusions has resulted in modeling as a major part of all such studies. As an example, models are constructed to investigate the optimum number of cooperative hunters that are required to kill a given prey. This takes into

consideration, among other parameters, the requirement of additional manpower in hunting a larger prey and the quantification of the amount of additional food that is so acquired. The observed number of hunters is then compared to the "optimum" number calculated from the model to decide if the behavior is "optimal" as would be expected if the evolutionary explanation was correct.

XIII.7 Strength of recent evolutionary scientific approaches

Conventional criticism of these evolutionary approaches by social scientists follows the pattern in economics. Qualitative arguments of the complexity of human system and cultural practices are invoked. Unfortunately, these are not logical proof of the validity or otherwise of the science. As has been mentioned repeatedly, all science is an approximate model of reality. The question is whether the complexity is quantitatively small enough and if the logic of the approaches is inherently strong enough to answer the limited questions asked. More important for the present analysis is whether this science can form a basis for societal decision making.

One statement can be made with absolute confidence. As in the case of economics, the functional forms employed in various models are purely empirical and have no formal strength. Apart from this limitation, a large fraction of the conclusions are based on statistical evidence. The limitations of drawing inferences from statistical evidence in the context of social sciences will not depend on whether the underlying scientific logic is evolutionary or not. These limitations will thus be discussed in the next chapter.

Returning to evolutionary explanations, as mentioned earlier, the primary strength of the evolution is the molecular picture. Problems in applying this to specific details of the evolution of a biological trait have been discussed. So the key question is the relevance of these limitations to the new efforts to extend evolutionary discussion to human behavior and culture. There are two basic aspects of Darwinian evolution. The first is random variation and the second is the limitation imposed on survival of the populations leading to selection. Considering the second aspect first, the major limitation of

applying the Darwinian approach to humans is very apparent. For whatever reason, even the early man, two million years ago managed to spread from the African grass lands to equatorial Java and temperate Beijing. The rapid emergence of the modern man from Africa started 70,000 years ago. Humanity managed to find all the continents and almost every ecological niche, from the tundra of Siberia to the deserts in 10-20,000 years. It is most important to realize that such rapid enhancement of the available ecological niche makes it extremely uncertain that the second part of the Darwinian scheme, filtration due to fitness is applicable. So when the population becomes too large to be sustained by a given ecological niche man finds another! This happens very slowly by genetic mutation in the case of other animals. In humans this happens very fast.

Consider the most significant genetic modification, lactose tolerance. If the population increase is largely unbound, the relative increase of the frequency of lactose tolerant individuals may represent a new reality. One where there is no competition between alternate phenotypes. The frequency has increased because more lactose tolerant individuals survived. But those without lactose tolerance also survive. One can even posit that the pastoral society is an additional ecological niche this time created by humans. In the case of the Polynesian migration, since the frequencies of genes in the founder population are not known, it becomes logically difficult to confirm selection. It is only an assumption that people with these traits survived and the others did not. Given the small founder population, there is a probability that this is only founder effect. However, the gene distributions are the strongest scientific results that can be assessed and they are important. Lack of immunity to smallpox certainly had a role to play in the history of America.

It is claimed that from a few hundred to two thousand genes show differences in their distribution among the population. However, no clear selection ideas have been identified with most. Uncontrolled desire for sugar and salt is almost universally cited as an adaptation in the early human environment that has turned mal-adaptive in the current culture of abundant availability, leading to obesity. Surprisingly, the genes for this trait have not been clearly identified. The strongest scientific support for evolutionary explanation of human

characteristics is for an earlier period of human history. The current conditions even in the so called underdeveloped world seem to offer very few choices for Darwinian evolution. Death and disease due to wars and man made calamities are regretfully significant and here chance plays a more important role in survival than genes.

However tenuous the importance of evolution in the gene distributions, the concept of Darwinian evolution of psychological and social traits appears to be even weaker. The issue is not the modular nature of the brain. That is clearly demonstrated by the study of human vision. There are many small modules in the process of converting the visual stimulus of the retina into the conscious experience of sight. The identification of these small models has enabled creation of many interesting visual illusions. There is clear evidence that many of these modules are formed by stimulation of the visual cortex. The problem is in identifying mate selection or tendency to homicide as modules in this sense. Stimulation and formation of neural networks corresponding to these modules by pruning of the random neural network is difficult to visualize.

Feynman describes in his usual funny way that what two individuals actually do when "counting" is not identical. He could read while counting but his friend could not. In a reversal of roles, he could not speak while counting but his friend could! Certainly the concepts of mate selection etc are more complex than counting. Can they be considered as genetically developed modules? The other significant point is that these are often conscious decisions. Distinguishing between a rational and conscious decision to select a wealthy bride from the influence of an "evolutionary" module is logically absurd. Similarly, it is relatively simple to accept that fear of snakes "could" be ingrained and any moving rope could act as the stimulus for its formation. But humans do have a capacity for overcoming this phobia as the snake charmers of India show quite apart from modern naturalists. Humans can and do demonstrate significant capability to grow out of the limitations of genetic inheritance

The possibilities of psychological evolved traits being relevant for current life as claimed by some of the research is further weakened

by the actual social norms and practices. While a hip to waist ratio may serve as a reasonable indicator for female attractiveness for men during tests, (and even this assumes that there has been no social conditioning) in many societies mate selection does not involve selection between uncovered females. Social norms and culture define the interactions leading to mate selection. The capability for preference of smell of males with low asymmetry in a controlled experiment may be statistically significant. However it is not very useful in a real world situation. The attraction of the scent should be strong enough in the presence of other smells ranging from scents to atmospheric pollutants. Surely women do not go round selecting mates by sniffing at them.

Then the same discipline of evolutionary psychology identifies other causes for mate selection while investigating bride wealth or polyandry. The interference between multiple possibilities is never discussed. More importantly many human behaviors are evolutionary absurdities, from celibacy to suicide. It is possible to create a possible scenario under which anorexia (an eating disorder characterized by refusal to maintain a healthy body weight and an obsessive fear of gaining weight) is an adaptation but clearly this is absurd. Similarly many possible explanations can be provided for female menopause including the propagation of genes and caring for grand children but there is no logical means of selecting among them.

The idea of "memetics" or considering human cultural and intellectual creations as exhibiting Darwinian evolution clearly contradicts the first percept of Darwinian adaptation identified earlier. The variability in these ideas or concepts is not random. The modification is a product of conscious decision by a human mind. In case of science or even religion, the modifications are initiated with a clear effort to appeal to the logical minds of the scientists in one case and the emotional susceptibilities of the population at large. Clearly as one moves from the study of gene frequencies to human culture the strength of Darwinian explanations diminishes.

Adaptiveness is metaphorically Darwinian. This to some extent parallels the use of Freudean and Marxist explanations in art and literature. There is always scope for intellectual satisfaction in

using these analyses but these cannot form the basis for any societal action plan. Perhaps the real "dangerous idea" is extrapolating Darwinian selection outside the domain for which it was originally developed since the extrapolation is not justified.

The above short description does not evaluate the strength of science in the individual studies of evolutionary explanations. This is best left to the individual scientists and their intellectual satisfaction. The present effort is a summing up to evaluate its role in supporting ideologies. The weakness identified is enough to rule out any confidence in its assistance to either the nature or the nurture camp. On the other hand the clear support of fundamental physics is a key argument against the ideological position of creationist science.

———————

XVI

SOCIAL SCIENCE : ANALYSES OBSERVATIONS AND IMPLICATIONS

XIV.1 Social sciences and humanities

Scientism is a charge most often leveled by religious believers on atheists and by the adherents of holistic principles as a response to reductionist efforts to reduce complex phenomena into simpler building blocks. Such philosophical arguments are as usual not relevant for the purpose of the present discussion. However another strand of scientism, though this is never labeled as such, is quite ubiquitous. Most demands for societal action, be it a simple case of calling for restrictions on violent electronic games, or provision of affirmative action programs, for handicapped individuals or historically oppressed sections of society are supported by quantitative social science. Analyzing the strength of such quantification constitutes a natural extension of the current investigation of the science behind ideologies of various types.

The use of quantitative social science is not confined to any specific ideology. In practice most ideologies, be they of right or left wing economics or of religious or secular persuasion seek its support. In most democratic societies, numerical support is often used as a

persuasive argument by groups to convert their special concerns into societal decisions. A short but fundamental analysis of the strength of these approaches is now attempted.

In the present effort, the word "social science" is employed to emphasize the use of the "scientific method" of quantification, as opposed to the descriptive approach implied by "humanities". The distinction implied here is not universal as the example of "political science" shows. In this example, "science" has been used to distinguish academic study from active politics. In the earlier chapters, words like rational, logical, objective and analytic were used in the context of science. While distinguishing social sciences and humanities it is important to underline the rational, logical and often objective nature of the analyses in the humanities. A descriptive language is doubtless the key feature but common human experience is the touch stone for this analysis. Logic and rational arguments are much older than the human efforts at quantification of knowledge. Socratic dialogues use extensive logical examination and Plato's analysis leading to the desirability and training of a philosopher king is extremely logical. So objective and rational investigation is always an integral part of humanities. Despite the scientific method and efforts at quantification, ideas and analysis in humanities are rarely converted to mathematical models in social sciences. The analytical models employed in economics and evolutionary explanations are an exception. But their limitations have been described in the previous chapters. The analysis and broad conclusions of humanities cannot be easily converted to the strict logic of mathematics.

One amusing example of the limitations of extrapolating logic is to look at amateur attempts at learning to play chess. The rules of chess are simple and a child gets to understand these easily. The first attempt a learner makes to improve his skill is to simply check the possible next moves. Sure enough, an amateur does not get too far in learning to play chess with this adherence to simple logic. A bit of thought would reveal the limitation of such extrapolation of logic. There are just too many possible moves and a correct procedure for learning or teaching chess is not based on such logic. Even computer chess does not follow this simple minded approach. The situation in many social sciences is similar. There are simple rules which are not

fixed by human convention as in chess but realized very easily from observation and experimentation. However, extending the experimental method that has been extremely successful in physical sciences to social sciences is an equally naïve approach. This has not prevented the desire to convert all of humanities into social science, ultimate example of this effort being experimental philosophy.

For a good example of experimental philosophy one can turn to the age old conundrum, whether it is morally acceptable to sacrifice one individual to save many. This problem has been turned into a thought experiment. A group of people are traveling on an uncontrolled trolley car running on rails, which is headed for an accident guaranteed to result in the death of all occupants. They can only be saved if this trolley car is diverted into a loop line killing a worker on that line. Thus one has scenario where many can be saved by killing one individual. In an alternative scenario, the group in the trolley can be saved by pushing a bystander in its path. Rationally and logically both are equivalent. In both cases, an individual is being sacrificed for greater good. Analyzing such situations has been part of ancient philosophical discussions. In the new science inspired approach, surveys are conducted to see if people in general find the two situations to be equivalent. Investigations show that far fewer people accept the second scenario. The experimental approach permits one to claim statistical significance to the claim that the two situations are not identical in human experience. Perhaps an evolutionary explanation can be sought. Thankfully, experimental philosophers explicitly concede that the investigations are to serve as a source for discussion and analyses rather than resolution, a position that most other social scientists do not concede.

XIV.2 Cross correlations and Arrow's theorem

The universally accepted feature of social science is the lack of fundamental mathematical theories. Thus, there is rarely any relationship between one empirically justified observation and another. This situation is unlike physics where, fundamental scientific theories emerged as relationships could be observed between observations in different domains. The lack of relationships between multiple empirical observations is equally a clear indicator for the impossibility

of a fundamental mathematical theory. In the context of deciding on possible societal implications, this situation is very important. One piece of evidence in for example, sociology may imply the suitability of a particular societal action while another say in economics could suggest precisely the opposite. The debates between nature and nurture become intractable for this reason. This also contributes to the hardening of ideological positions (either of nature or of nurture) and questioning the empirical evidence supporting the other side. The inevitability of irresolvable conflicts becomes apparent when Arrow's theorem is analyzed in a novel way.

Arrow's theorem is usually used to evaluate and compare election procedures. It is also an important component of economic thought and is used for logical analysis of economics of social welfare. The theorem questions the possibility of converting a set of individual choices to a rational social decision. Usually this is discussed using an example of a set of voters electing an individual for a political office in a democracy. Each voter has a choice of a specific candidate for the position. In a practical democracy, there are many possible ways of selecting. The most usual is election of the person with the highest support among all candidates. Arrow's theorem asks is if there is a possible logical or rational conversion from a set of individual choices to a justified rational collective decision and logically proves that there is none.

It is this mathematical rigor that points out to the possibility of extending the idea outside this limited range. As usual in the present approach, we look at this logical deduction as a limitation on the extrapolation in the spirit of Gödel's theorem encountered in an earlier chapter. We thus equate the various choices to a set of rationally deduced empirical observations. In the absence of any relationships between the individual empirical results, Arrow's theorem serves as a warning that a comprehensive integration of the empirical results is "not always possible". Thus one notices another logical limitation for the applicability social science research. The theorem makes some assumptions about the rational individual decisions. For example it assumes that the rational preference of possibility "a" over "b" is not affected by the introduction of a new possibility "c". It is empirically known that these assumptions are not followed by practical human

economic choices. For example a customer may prefer black when offered a choice between white and black but choose white from among white, brown and black. Thus, the above is a not a proof but as in the earlier use of deterministic chaos, a warning that problems emerge even in extremely simplified theoretical situations.

XIV.3 Inherent limits of quantification in social sciences

As has been discussed in earlier chapters, statistical correlation in the absence of fundamental theories is the weakest form of quantification. Unlike even empirical functional descriptions, there is no concept of an independent variable that can be controlled and repeatedly set to identical values for acquiring data. The limitations of employing continuous variables and mathematical analysis even in the most mathematical social science (economics) have already been discussed. In social sciences, in the absence of analytic mathematical functions only two approaches are employed. One is to determine the correlation coefficient and the second, to determine the statistical significance. The earlier discussion pointed out the limitations of these approaches. The high probability for the occurrence of the opposite of a proposition proved to be statistically significant was discussed. As was observed, "the average height of man is larger than that of a woman" may be verified to be statistically significant at 5% level of confidence but this does not ensure the probability of a randomly selected woman being taller than a randomly selected man being 5%. The tendency of short runs of order that appears in randomness as exemplified by the continuous series of heads or tails in the tossing of a fair coin showed that even when a clear statistical confirmation of a preposition is obtained, the possibility exists that this is purely a random event. Similarly, correlation can be obtained but may represent nothing but a transient deviation from the mean that could be purely random. Finally errors tend to increase as the absolute magnitude of the probability decreases due to base line fallacy.

Such limitations are significantly more important in social sciences than the other areas considered so far. Some of these problems can be reduced by proper choice of the population for testing and

careful use of statistics. Obviously in a small summary as is being attempted here, it is not possible to go into these technical details. A few general features common to all social science approaches will however be considered.

Most concepts employed in social science are descriptive. Most concepts used in physics are also often described using words. But the words, force or pressure for example are defined in a specific way and related to a procedure of measurement. In an earlier description of gas, pressure was defined as the momentum transferred to the container by the molecules of gas. For completeness, it may be pertinent to admit that, in recent years with the introduction of quantum mechanics, purely mathematical entities that have no real operational reality like the wave function have entered the language of physics. However, even these purely mathematical quantities are converted to physically measured quantities mathematically. For example, the magnitude of the wave function gives the probability of observation of an experimental result.

Unlike in physics, the relationship of concepts and quantified variables in social sciences to experience is tenuous. Some relationship is obviously necessary and exists since there is enough communication to lead to earnest and often heated discussions. However no universally accepted operational meaning emerges in social sciences in contrast to natural sciences. It was seen that even the most widely used quantified concept, "money" is quite complex. In the modern society money is not merely the currency notes available in circulation. For that matter a currency note is probably not really "money" since it only "promises to pay the bearer a sum of …." Defining monetary value for large transactions on the stock market is problematic. It may be news worthy to prepare a list of billionaires but they mostly own shares in the stock market. If even a fraction of these shares is actually sold, they would yield only a fraction of the claimed wealth. If the billionaire cannot exchange the present set of goods and services held by him for another without significant loss, does it make any sense to count his billions in the first place?

Similarly a psychologist may certify the "sanity" of a prisoner under trial but would be incapable of distinguishing between the

"sanity" of two individuals. Thus one has a concept that cannot be used for small deviations though it has an approximate value for large deviations. Even in the case of a term like intelligent coefficient, a parameter that is defined operationally, based on the response of an individual to a standardized test procedure, the relevance and utility of the term are extremely contentious.

Most descriptive terms would in reality be a continuous spectrum of values and opinions. It is usual to label an individual as atheist or religious but as anyone knows this dichotomy is a forced simplification of complex reality. A further limitation of this qualitative description is the variation of the response with changes in description. A small change in the wording of the question results in large change in the responses elicited. For example, in the United States, public support for expenditure for "poor" gets more support than for "welfare of poor". Such observations are most surprising if only a linguistic analysis of the changes is made. All this makes formulating criteria for defining the "sample", a group of people with identical capabilities difficult since the parameters describing the individual are once again descriptive and vague. While it is quite common to see "statistically significant" evidence that religious people are happier, live longer and recover from diseases faster, the selection of the control group is even in principle quite difficult.

A similar problem of converting from a continuous variable to a forced dichotomy was earlier discussed in the case of medicine. It was pointed out that the consequence of aspirin should have been a continuous decrease in intensity of the pain. In the interest of making a statistical comparison between use and non use of the medicine, the complexity is simplified into a prior definition of cure and no-cure. At least in medicine, familiar quantitative measures such as pulse, temperature and blood pressure, supplemented by an ever increasing range of biochemical analyses and other investigative tools are available to help in diagnosis and in selecting "reasonably well defined control sample". This advantage is the consequence of accepting the contingency of physics and chemistry. "Alternate medicine" does not have this feature either. Social sciences do not in general have such additional parameters and tests for selecting the control group. Another issue that emerges in extrapolating social science conclusions is the

effect on the parent population. Small and significant phenomena can be logically justified or empirically observed. However, the magnitude will not be significant. Extrapolating to larger magnitudes often means change in the original population.

The most common and simplest instance of this is the adulation bestowed on an act of large philanthropy and the call for emulating this. If philanthropy becomes significant it will alter significantly the economy that has created the rich individuals. Gandhi envisaged an economy where the rich were trustees rather than owners of wealth. This may appear attractive but extrapolation from current philanthropy levels is naïve. If every rich man decides to donate, there are no buyers of the wealth. Also the section of the economy providing luxury goods and services will become jobless. Contrary to socialist emotions, "ostentatious" expenditure of the rich cannot be entirely diverted to charity. This is true of religion too. Notwithstanding the merit in these life choices, a Christian society cannot consist entirely of nuns and lay brothers or a Buddhist one entirely of monks. Giving alms does not merely need the rich and the poor. It also assumes that all the rich will not simultaneously donate.

A final problem with extrapolating from social science results is the problem of conscious decisions. Consciousness as a limitation on evolutionary explanations has already been discussed. In general this is true even if the social science result has a non evolutionary explanation. The above two limitations are most commonly ignored when economic logic is used to provide advice in social context. As mentioned earlier, in recent years, logic of the free markets has been extended to various social situations ranging from selection of sexual partners to human dishonesty. The arguments and logic in various popular expositions are most persuasive till one recognizes the two limitations outlined above (in addition to the limitations of the logic outlined in the earlier chapter).

XIV.4 Analyzing the utility : A few examples

The above has been a vague and general description of the limitations of the social science research in offering guidance for societal decisions. Since the area is so vast it will not be possible to

review all of it, show how these limitations influence virtually every aspect of the research and caution against extrapolating the evidence into strong convictions. However, as in the earlier discussions concerning economics or climate change, the essence can be justified by few specific examples.

When utilization of social science research in guiding societal norms and laws is evaluated, several possibilities exist. Sometimes the issues are contested and there is no clear consensus in the society. This is by far the most common reality. This does not mean that the opposing positions are equally valid. This had become evident in the climate change discussion. The conclusions clearly showed that while confidence in our ability to confirm various claims of climate change is variable, blanket opposition to global warming is scientifically much weaker. Sometimes the support for a weak scientific position can be quite strong. This for example is the case with alternate medicine or the scare about mobile phones.

The second possibility is a strong consensus, the societal norms reflect the opinion of social science but there may not be legal restrictions. Finally the strong consensus results in a legal framework and the society feels it justified to use coercion and force the members to follow the norms.

Single examples have been taken up for each of the three positions highlighted above namely, lack of consensus, an effective consensus and finally legal coercion. Relatively non-controversial examples have been chosen. The goal is to highlight the various limitations of generic social science conclusions as outlined in the previous discussion in these cases.

A detailed discussion of available knowledge to asses the strength of the science in each example is not attempted. Unlike in earlier cases like climate change, there is no common fundamental theory that can help understand the scientific strength of the consensus or the lack of it. Consequently, such detail is really not attractive for the present effort on understanding how well we know. We simply concentrate on the strength of the empirical observations in the light of our earlier discussions.

The first relatively benign example is the influence of violence in games particularly video games on children. The availability of video technology with its graphic display of violence has energized older concerns regarding the use of toy guns and the implication of playing with gender stereotyped toys during childhood on the roles boys and girls are expected to play in later life. There is no dearth of empirical surveys that claim statistically significant correlation between playing violent games and tendency to violent behavior in later life.

As mentioned earlier we will not look in detail at the scientific basis for the link between playing violent games and violent actions in the overall context of human development. We simply analyze the possible strength of any such empirical evidence. To begin with defining violence in the game or other activity accepted by society is very vague. Specifying "violent activity" in the adult is also quite problematic. Thus categorization of games or adults as "violent" or as "not violent" is an artificially created dichotomy.

In addition to these descriptive limitations, not everyone who plays "violent" games, video or otherwise as a child is prone to violent activity as an adult. Despite the huge alarm that is raised, most people are not violent most of the time. Violent actions for most people are simply an aberration and a rarity. Thus one has a very small number with the associated errors as discussed with base line fallacy. The accuracy of the statistical data is bound to be quite low.

Many other factors ranging from lack of education, dysfunctional families, weak social support services, alcoholism etc are better indicates of violent behavior even among this small fraction. Even without any deep understanding of the science, these factors make much more sense as predictors for violent behavior. Thus it would have been quite difficult to ensure that these factors do not contribute reducing the strength of the evidence even further.

Despite the weakness of the evidence, demands are routinely made not only demanding such action but also questioning many long standing cultural practices ranging from carrying the kripan (a small ceremonial sword) by orthodox Sikh males to reading the many

descriptions of violence in Bible classes. This ignores the obvious scientific factor that if such minor activities are as important as is made out, violence should be much more common in groups or societies with these practices. The absence of empirical studies supporting this is simply brushed under the rug. This demonstrates that the demands have effectively formed an ideology by extrapolating from very weak science.

Finally, one comes to the usual dilemma that plagues each and every societal action. Assuming that a small fraction of children are at risk, the cost of creating a social infrastructure for examination and certification of the video games and punishing the transgressors has to be assessed and compared to other possible methods for correcting violent individuals.

As a second example of a consensus that has emerged in societal values because of the findings of social science research we can consider the requirement of gender neutrality in language. The earlier practice of using masculine words has been largely replaced by gender neutral terminology. A common example is the replacement of chairman by chairperson or a batsman by batter (which also is a mix of spices and flour coating something to be fried, or the mixed ingredients for baking a cake). There is every possibility of finding research that shows that individuals who do not use gender neutral language are also statistically prone to other acts of discrimination based on gender. A similar claim may be made for people using language ignoring racial or other group sensibilities.

To be sure support for this modification of language exists without any great expectation that merely the modification of the language will contribute to racial or gender equality. The consensus has not been converted into a legal requirement. However, this "sensitivity" has been extrapolated to absurd lengths as in objecting to simple "Merry Christmas" greetings. It should be most obvious to the supporters of politically correct vocabulary that a simple conscious decision to conform without changing the underlying discriminating tendencies is very probable. This is an example of how consciousness can derail both empirical data of social science research and societal norms. More interestingly, the supporters of this "training" would

scream that such societal constraints and conditioning is counter productive and inhuman in other contexts such as sexual orientation.

One interesting example derails this entire linkage between language use and discrimination. One is familiar with the absence of a neutral gender in languages such as French or Hindi with inanimate objects being labeled masculine or feminine and the appropriate verbs being employed. There is one language, Telugu where the masculine words are separated and the feminine and neutral gender words share the same verb forms. Obviously if the use of language was either a significant contributor or indicator of gender or other discrimination, such large scale differences should be reflected in the social conditions whereas they are not.

Finally, consider one other issue where there is not only a contemporary societal consensus but an effective legal mechanism to punish the transgressors. The motivation for both the consensus and the legal injunction is social science research. Corporal punishment particularly of children was never proscribed in any society. "Spare the rod and spoil the child" was the refrain of most societies and cultures till recently. In a country like India, the norm has changed from acceptance to media outrage and legal punishments in less than 25 years. A demand for proof of negative consequences of corporal punishment would be considered lunacy and news reports of attempted suicides appear to make this a justified response.

A curious but ignored point is the strong correlation between the negative attitudes towards corporal punishment in the society at large and the increased incidence of the drastic actions by the younger generation. When such punishment was the norm and widespread, attempted suicides were unknown. Today, when the norm is respected, there are many attempts at suicide due to "humiliation" associated with not only corporal punishment but even a public comment. As usual, conscious awareness has a major consequence in social issues. This is also an example of the change in parent population brought about by the social changes themselves.

The negative effects of corporal punishment and its ineffectiveness as an agent for correction are routinely highlighted.

More "acceptable" psychological methods are much praised but curiously these methods are proposed, praised and forgotten within very short time. This rapid circulation of approaches shows that they are at least as ineffective in correcting undesired behavior as the much maligned corporal punishment. Attributing the lamentations of modern parents to merely their own conservatism or incapacity is merely acceptance of a social norm and not real empirical justification for the efficacy of the alternates.

Actually the problem is much more subtle. This subtlety is lost when the consequences of corporal punishment are reduced from a continuous variation to a dichotomy. The influence of any societal norm on the members of the society varies. Thus some individuals may not be affected seriously by corporal punishment but a small fraction who are more sensitive may be. The relative size of these segments will depend on the society. The parent population continuously changes. It is always emotionally appealing to demand societal norms which are absolutely "safe". The support for this precautionary principle increases with economic progress of the society. However, as was mentioned in the discussions on medicine and environment, implementing the precautionary principle is beyond human capability.

The above criticism must not be seen as a justification of any of the positions. No justification of insensitive language, violent games or corporal punishment is intended. The above is merely an examination of the degree to which these societal norms or legal sanctions are scientific. It is to highlight the extent to which simple, largely accepted societal norms are collective emotional positions despite attempts to dress the arguments in the language of science.

The reason these simple examples have been selected is to demonstrate how even in the simplest cases, social science research, while providing unquestioned intellectual challenge has severe limitations when used as justification for societal action. When examined logically, the support for these positions is based on social awareness and some kind of gestalt reasoning and not rational arguments. This important role of social science is discussed in the next section.

XIV.5 Analyses and their implications

At the start of the chapter, the analytic tradition in humanities was mentioned. Philosophy, literature, criticism and knowledge influence human decision making even if the mechanisms and processes are not clearly understood. The best example in literature is the great speech of Mark Antony in Shakespeare's Julius Caesar. The transformation in the perception of the listeners from considering Caesar to be an ambitious villain to a hero may be only a literary expression but one that resonates with the reader. Many claim that the support for the American Civil War was a consequence of the description of slavery in "Uncle Tom's Cabin". While many attempts at psychological analysis of these processes are made, such rational examination will once again be subject to the limitations that have been discussed in the last few pages. Clearly conclusions of research in social sciences can influence the human societal action whether they employ quantitative terminology or not. This does not mean empirical observations are unnecessary or even irrelevant. Merely that they do not result in an unambiguous resolution of the issue.

Each issue that necessitates a societal action would always be helped by the availability of empirical data. However, since empirical science fails to provide rational decision to choose, analysis based in the tradition of humanities should be seen as at least as creditable as the respected scientific quantification and not discredited or dismissed because of the poorly understood mechanisms. The success of the successful manager with the "gut" feeling but no MBA degree in running business enterprises and of the homely philosopher without a degree in psychiatry as an advisor for self improvement show the importance of these approaches. In a range of activities ranging from appreciation of art and grading of wines to teaching, the role of quantitative measures has been demonstrated time and again to be useless in practice. The many efforts to quantify scientific research through numerical indices like the impact factor fail for similar reasons.

Thus, overuse of mathematics is often counter productive and is an effort at covering up ignorance and pretending expertise. Once due weight is given to what has been termed above as the gestalt

consequence of social science knowledge or of humanities as they are better labeled, one interesting heretical thought emerges. How different is such influence of descriptive humanities from religious practice and thought, even if religious practice proceeds from rationally indefensible customs and theology? The ancient and controversial idea of "two magesteria", the regimes of scientific investigation of the natural world and moral guidance of the religions has been revived in the recent past by Stephen Gould. Perhaps the real value of religion and philosophy lay in providing the individual with some capability to handle the dilemmas exposed by scientific enquiry thus making the two magesteria more intricately linked like the duality in quantum mechanics. This however is a thought that will be explored in a sequel volume.

Summary

Limits On Utilizing Science

Application of science for human betterment has been the progressive call ever since the grandeur of Newton's achievements was realized. Physicists have been able to provide a precise understanding of nature and use the knowledge for improving "Human Development". Optimism permeates that all human problems and dilemmas are resolvable by a similar process if only we think about them. Rare is a Feynman who admits that scientists actually fail to make any progress when they tackle large human problems, let alone solve them. While the desire to intellectually accomplish as much as physics in every other science has been a great motivation for progress of every other science, an equally strong desire to assist human life cannot be overlooked. Philosophers and humanists since antiquity had been equally driven by an empathy to the human cause. The story of Prometheus, who according to Greek mythology stole fire from the gods for humans and was punished for it may have been created as a warning but the desire for application of human knowledge always existed. Early attempts did not create an objective rationally justified mechanism nor was that attempted. While Socratic dialogues represent a great effort at rationally discussing human problems, there was no attempt to model this on the line of geometry and end the discussion

with a QED "quod erat demonstrandum" (which was to be demonstrated). The philosophical discussions since Socrates have tried to improve the precision in language and description in an attempt to improve the understanding. While physicists have largely given up the hope of communicating quantum theory in everyday language a philosopher continues to harbor the fond hope that precision in language can help him in understand "everything".

On the other hand, there have been the efforts to employ every possible mathematical trick in an effort to improve the scientific understanding and more importantly utilize the quantification as the justification for influencing societal decisions in line with these quantified conclusions. The earlier chapters show the limits of this quantification enterprise in matters of societal concern. The true appreciation of the strengths and weaknesses of mathematics and science was provided in the first two parts. In order to assess their impact on issues of societal concern, we had employed the known conundrums and complexities of medical science and health as the paradigm and cast these issues in the framework of the doctor's dilemma. As in any area of human activity, quantification and logical exploration do help medicine in many ways. Newer and better science can and does provide better options for resolving the doctor's dilemma in specific cases. However, we have seen that extrapolating this success into a philosophy or ideology, expecting the resolution of all problems is futile. Decisions have to be made by the physician, the individual or the family while being subject to a terrible dilemma. The process whereby the physician integrates his experience to come to a specific advice for a specific situation cannot be rationally analyzed. As one moves away from decisions concerning the individual to societal decisions, medical decisions become even more problematic.

The increasing distance from fundamental physics is one important component of this increasing uncertainty. The other is the emergence of complexity in the systems leading to the impossibility of rational or logical deductions. This was first encountered in the discussion of the reliability of a MOSFET. Similarly, in discussing climate change, we found that the physical processes are well understood but the complexity of the system defeats confirmation of predictions. This limits human abilities to tackle the environmental

problems. In continuation of the earlier discussion of medical issues, this was dubbed the environmentalist's dilemma.

With problems in economics, psychology and other social sciences, a link to fundamental physics is almost nonexistent. Even evolutionary explanations can only provide a limited linkage and cannot be conclusively applied to specific problems. The conclusions drawn from purely statistical analysis of data are extremely fragile and can be applied to local problems but have no strength to be used as support for societal decisions in a general way.

The basic fragility of statistical analysis of data with respect to the probability description of significance, the base line fallacy and random possibilities in short runs of data have been analyzed earlier. Two additional complications were highlighted in the context of social sciences. One is the ancient issue of consciousness. The efforts of all the philosophers in the world since antiquity have not helped in formulating a consensus on what consciousness is. For our purpose however the problem is more modest. What are the consequences of conscious knowledge on conclusions drawn from observations of the society? This issue was mentioned when evolutionary explanations of human behavior were discussed and again in the previous chapter for other social issues. It is not only a question of distinguishing between a conscious decision and a decision caused by brain module that emerged as an evolutionary adaptation. It is a simple truism that to a large extent humans by conscious effort can and do change their habits and behaviors. One usually sees the positive side of such abilities. Certainly exposure to the inherent cruelty of slavery or racial discrimination could help in changing human habits and overcoming traditions. At the same time, the banality of evil associated with ordinary German citizens of the Third Reich is a warning of the reverse possibility.

In addition to the conscious decisions of individuals, as the awareness of the issues spreads in the population, the population itself changes and the earlier scientific conclusions may not be applicable. This is most commonly accepted in discussions on the stock market. The indices change with perceptions and even rumors leading to closed loops. This is true in general for all social issues. The role of the

general acceptance of norms was also discussed in relation to corporal punishment in the previous chapter.

The issue of non-linearity and complexity in systems leading to large fluctuations was encountered with the environment and is common experience in the economic activity. However, death can also be considered a large and irreversible fluctuation of a complex living system. Such fluctuations are in more common language termed as accidents. So not merely death in an accident but death itself can be considered as an accident. In the case of MOSFET, as we moved away from fundamental physics and empirical relationships one ended with failure statistics. Once again these share the essential feature of an accident namely, inevitable, unpredictable failure of complex and tightly coupled systems. In the case of MOSFETs, the ability to test a large number of units to destruction enables reasonable estimate of failure probabilities. If one considers a passenger airliner, another complex system with an extremely low failure rate, confidence in low failure rate emerges from detailed analysis of the failure of individual components. In view of the inherent variability of living organisms, despite the best efforts, death of an individual living organism becomes less predictable. Sudden death and miraculous cures are experienced by every physician. In systems without the facility of statistical analysis over large subunits, there is no rational way of even estimating the failure rates. This does not preclude claims of managing failures and accidents to avoid open admission of the limitations of human capabilities. These approaches have more in common with the desire to use non traditional medical approaches in cases of fatal diseases than the practitioners would care to admit.

The most interesting aspect of the discussion of the previous chapters is how easily, and without understanding the academic research of the subjects, it is possible to form clear ideas how well these are known and how far they can form a guide for personal and societal action. Thus one returns after a serious quest for Newtonian science outside physics to the earlier Socratic and philosophic aim of integrating specific knowledge into wisdom. But this is ultimately an individual psychological process. Integrating the science for societal decision making remains problematic and the next part explores the only possible resolution.

Part Four

Resolution Of Dilemmas

Richard Feynman approvingly quotes this advice offered by a Buddhist monk. “Every man is provided with a key to the gates of heaven. The same key opens the gates of hell”. This is an example of a typical dilemma. Obviously the key is required but the key in itself cannot distinguish between the gates of heaven and hell. The earlier chapters logically demonstrate the presence of this dilemma in every human scientific endeavor. Science like the key is required but one has to be clear about what it can and cannot deliver.

It is possible to ignore the essential problem and dream about scientific resolution of all issues including moral queries. It is also possible to extrapolate current scientific success either to a heaven of eternal life on earth or the hell of environmental chaos. Resolving tangled ideas is a real problem and one needs a holism that enables a reasonable and practical approach for resolution of dilemmas. Integrating the various dimensions of the dilemma really constitutes a proper domain for holism as a counter to unsupported extrapolations of science. Sadly, the term has been often employed as a battering ram for questioning scientific research for emotional reasons or for ignoring science all together.

There is a long history of advice that seeks to enable the individual to resolve the dilemma. Unlike putative science advice of modern psychology or other sciences, these analyses offer advice and hope for the emergence of individual wisdom. The range is really vast, from the ancient Socratic advice of the golden mean as the desired goal to the famous statement made by Neils Bohr in a different context. "The opposite of a small truth is a lie, the opposite of a big truth is another big truth". These are nothing but different facets of the ancient philosophical recognition of the duality that plagues man, the duality between, happiness and sorrow, life and death, truth and beauty and so on. There is an explicit acceptance of the impossibility of rationally resolving these dualities.

The resolution of dilemmas is required in two different practical situations. The first involves a decision by an individual when the consequences of the action are purely personal. The second is a communal decision where the opinions of most if not all members of society have to be accommodated. The consequences are societal in nature. Resolving the dilemmas or cutting the Gordian knot in these two situations is the issue that we shall discuss next.

Having distilled an essence of various disciplines, ignoring their individual complexities, the next part takes an even bolder step of trying to outline mechanisms for practical resolution of these issues which have been identified during the previous few chapters as being without rational resolution. These personal philosophical ramblings are being presented without inhibition though they are not novel and often repeat ancient wisdom.

———————

XV
WHAT IS BEST IN PRACTICE

XV.1 Cutting the Gordian knot : Individual and society

As mentioned earlier, resolution of human dilemmas is required in two different practical situations. First, when an individual has to take a decision. The individual may act as the supreme or final decision making authority because the consequences of the action are purely personal. Several examples have been cited during the course of the analysis over the last few chapters. Decisions regarding personal health, economics and social choice are justifiably personal since the consequences are borne by the individual. The second practical scenario implies a communal decision where the opinions of most if not all members of society have to be accommodated. Several examples were discussed earlier ranging from the decisions regarding provision of immunizations to the current universal concern about climate change. Science can illuminate each of these problems but cannot provide a rational resolution. Acceptance of the scientific research could vary from one individual to another. A resolution of the dilemma at the community level would thus require some kind of holistic integration of all these opinions.

Apart from the above two situations, there are situations where the decision is taken by one individual while the consequences are borne by others. Examples are the guardian or parent taking decisions on behalf of children or the incapacitated, the head of a commercial undertaking who is explicitly provided the power to take the final call and a dictator (kings in the earlier period) who assumed absolute power. In the case of guardians or managers, the powers vested in them and the limits of these powers are societal decisions in their own right. Absolute monarchs and dictators have assumed total powers. The modern democratically elected leader may also claim that the "buck stops here" but in reality is different from a dictator in that he attempts to provide an integration of the societal opinions.

As mentioned earlier, science can and does provide alternates and options. It is quite possible that this resolves the dilemma to a triviality. If there are two alternate medical procedures and one has a significantly lower side effects, it does not need a Nobel Prize to decide to accept the better option. The problem is with the identification of principles for action when the situation is not so obvious. Hopefully the discussion of the previous parts enables one to accurately gauze the relative values of the various options available in a given situation without being bamboozled by claims of expertise. Unfortunately, unlike the assumption of absolute powers by dictators, assumption of expertise has largely gone unchallenged. The goal of the earlier chapters is to convince the reader that while expertise to actively participate in the academic activity is beyond his grasp, it does not require much effort to learn how well we know it and decide his personal course of action.

What could be a rational resolution of the dilemma? If the approach of the previous chapters is strictly applied, one comes to an inescapable conclusion that every action is rationally a trial. Actually, "try it and see" is a common accepted scientific procedure. When the logic of established science does not result in a clear decision regarding experimental conditions for a proposed scientific investigation, one just tries. The important issue is the realization that the consequences are not predicted to a reasonable certainty and there is a clear possibility of the trial leading to an error or failure. Thus one accepts the trial and error in a spirit of humility, with the realization that this is being

done for want of a better option. Thus the seasoned physician will call clinical diagnosis and treatment more of an art than a science. This resolution is the reason Budian's Donkey described earlier is only a story told in philosophical circles. The most important point is that "doing nothing" is itself one of the options available for trial and not anything special. There is no special validity for that particular option.

To draw a conclusion at the end of two hundred pages of describing science that one is really free to do as one pleases may appear quite silly and seriously question the utility of the earlier discussion. As far as personal decisions are concerned, this may be acceptable except to those who ideologically refuse to give the choice to the individual. Even an individual may not accept the logic of trial and error and take a decision in tune with his personal ideology. If the consequences are borne by the individual personally there cannot be any alternative to accepting this as long as individual freedom is accepted. In the case of a commercial undertaking, one assumes that contrary to recent practice, commercial losses will not be nationalized and borne by the state and society so the manager can take decision based on ideology.

There are severe limitations of human psychological capability in accepting and acting in this uncertain situation and some of these will be explored in the next chapter. This does not in any way alter the basic argument of accepting the individual's decision in practice. There is one advantage of accepting decisions based on an ideological position. Most conveniently, the failures will never be attributed to the extrapolation of ideological positions but only to the complexity that exists in real situations. The dictum that victory is the general's greatness and defeat the soldier's weakness is applied in various forms.

Thus one never comes across a patient using non traditional medicine and counting the failure as evidence against the system. It is mostly attributed to bad luck or different personality or complexity etc. But any success is unhesitatingly attributed not to luck or chance but to the greatness of the method. Similar is the response of the economic ideologue whether of the free market or Marxist variety. Failures are attributed to not enough ideology, not to failure of

ideology. The failures in actual practice of several societal decisions based on "current social science research" in issues ranging from prison reforms to education are similarly never credited to the questionable science.

XV.2 Resolving societal dilemmas

Integrating diverse and often contradictory assessments is necessary for individual decisions also. For example, studies extolling the benefit of one food or other in preventing a specific disease are numerous. Surely, all these cannot be simultaneously consumed. So further research cannot realistically resolve the issue. "Balanced diet" will remain the best option. Similarly all recommendations of suitable exercise or physiotherapy cannot be simultaneously implemented (even if there is science to backup these and not merely commercial advertisement). However, there are many issues, ranging from what if any has to be done about climate change to the advisability of adding vaccination for cervical cancer to the national immunization program where each individual member of the society has an independent choice but the society at large can only take one final decision.

Integrating these opinions will be the responsibility of the "experts" or rulers in some societies but in the so called open societies, this integration will take the form of open canvassing for and against the specific cause. In the cacophony of open societies, strong vocal support is articulated for one's own position while the scientific basis of the opposite position is questioned. Ultimately, as with the case of most social science, the continuous variable representing the degree of support is converted into a statistically significant binary assessment of for and against.

This is quite easily seen with the climate change program. An earlier chapter tried to make a balanced analysis but that may not appear balanced to everyone! The variable strengths of various aspects of the science of climate is often changed into a single point agenda that claims the onset of a global catastrophe and further a blanket opposition to all forms of carbon emissions. The bipolar conclusion places one either as a supporter or opponent of an action plan for climate change mitigation.

Obviously, one has not only a limitation that the individuals may differ in their final assessment but also the degree of the support or opposition. A further limitation of this all out effort for garnering support to one's own position is the agglomeration of individuals who come to the same final position but from another starting point or another reason altogether. For example, it is possible to be simply a skeptic as far as climate change is concerned but oppose a given project based on the support for preservation of wild life or the life style of local population. As another example, while the terms believer in God and atheist are used as descriptions of individuals for opinion surveys, in practice this covers a wide range of personal positions. If one juxtaposes the degree of belief in God against support for abortion, one could expect that the most devout are the most opposed to abortions. But this will not only vary with the religion, devout Buddhists and Hindus may not show the same level of opposition, but even a hard core atheist may be extremely opposed to abortion because of commitment to human rights. Many extreme supporters of sacredness of life are atheists. Many of the supporters of PETA are atheists. Thus, in the context of a societal decision, the situation is extremely complex.

The situation is so complex that there is little one can do except to accept democratic decision making as inevitable for choice under ignorance and equivalent to trial and error. It is significant that some of the earliest political philosophers creating the first constitutional democratic system clearly described the attempt as a trial and error approach to finding societal rules in the spirit of the scientific experimentation. But for such a philosophy, the provision to amend the constitution as per current requirements makes no sense.

XV.3 Mechanics of democratic decisions

The common procedure in discussing democracies and constitutions is to take a historical view and perhaps pay homage to the Greek democracy. In these days of political correctness one could perhaps make the tribal societies and councils of elders in various cultures share the glory. One can also provide a comparative advantage to democracy by either empirical data or simply quote Churchill. "Democracy is the worst form of government except all the others

that have been tried". However, if one considers that the reason for having a democratic decision is simply the impossibility of any other mechanism supported by empirical, rational or scientific evidence, one can avoid much of this historical research. One can also ask very relevant and fundamental questions about the nature of some of the mechanisms of democratic decision making.

As mentioned earlier, Arrow's theorem has been often discussed in the context of comparing the various electoral practices. This theorem is usually employed to discuss the selection of one candidate among many by a large number of electors. If there are say N candidates who are available, each voter can form his own priority list of the candidates. Assuming that all these priority lists are available and are rational, (for example, the order among the existing candidates does not change if other candidates are added or removed) Arrow's theorem makes a mathematical analysis of whether any rationally justified common list of priority can emerge. However, the key issue of why an individual must choose a representative and how he can do so is not addressed. The question that we seek to address here is why we need to have a representative in the first place.

Switzerland offers a working example of a democratic nation where many decisions and individual laws are decided by referendum or direct polling of the voters. With modern electronic connectivity, establishing this type of decision making is quite easy. Since democratic decision making necessarily means taking decisions in ignorance, a direct procedure, as practiced by the Greeks appears appealing. However, the Greeks actually prevented both women and slaves from participating.

As mentioned earlier, individual preferences are likely to be cross correlated. If a voter supports a particular candidate for political power and decision making authority because he is most likely to support active measures to mitigate global warming, the candidate may also strongly support abortion rights which the voter may not approve. It is not clear how the voter practically sums all the positions of the candidates on various issues and supports one candidate even though this "holistic choice" is being taken in every democratic decision. On the other hand, if the voter has direct participation, he

can support one and oppose the other issue. The voter may be a minority in both and both decisions may go against him, but that presumably could prompt the voter to canvass and convince his fellow citizens, at least a theoretical possibility.

Contrast this advantage of the direct participatory or referendum approach to selecting a representative. The voter has to make a list of all the issues important to him personally and know the response to each of these issues from all available candidates. Even this would only enable him to make a huge set of lists ordering the preferable candidate for each of these issues. It still will not enable one to logically select one candidate. In other words, the lists which are the basis for discussions based on Arrows theorem are prepared by a non rational procedure even if the theorem itself is an example of perfect logic.

In reality, most societies do not use referendums extensively. Even Switzerland operates with the usual elected representatives for many decisions. It is the societies that do not have a tradition of democracy which call for plebiscites. This overwhelming preference for representatives over participation is actually very sensible. There are three logical reasons for this. The first is the obvious and most cited problem of organizing the referendums. There is a clear economic cost. However, electronic communication technology has enabled the possibility of instant surveys with good statistical basis. These already do influence activities of both the legislature and regretfully the judiciary. However, such procedures which bypass the representative tend to overlook two other limitations. First is the lowered possibility of compromises. When an issue is strongly polarized, a representative being a compromise over many issues could take a moderate approach. Secondly, over reliance on the direct participatory decision making permits rapid changes in procedures with equally rapid reversal of decisions, leading to continuous disturbances and lack of stability. The demand for a right to recall the representative is also an effort at rapid change, ignoring the next election as an existing albeit delayed right to recall.

The first of these limitations can be (to some extent) limited by the great idea of a two third majority. A reversal of a referendum decision made with a two thirds majority will need at least a majority

of the former adherents to change their position, (if none of the original opponents, less than one third to begin with, have not changed their preference). While accepting this two third majority rule for referendums and direct decisions reduces the possibility of rapid fluctuations in decisions, the legal system in the democratic societies works as another mechanism for ensuring that rapid changes in the opinions do not create uncertainty in the system.

Describing the legal system in modern democratic societies in such a language immediately provides an analogy with automatic feedback control systems. The most common example of this is a temperature controller or a thermostat. In any real physical system, generation of heat such as by a heater or conveying of heat from one place to another through a blower or flow of water is not instantaneous. Also, there are variations in the amount of heat required at different locations. Very precise control, required in advanced systems, is accomplished by using just one or two time delays in the control process. Thus, controlling a complex system is better achieved by a simplified and delayed decision making rather than accommodating each local requirement. The use of delay in the political system of decision making is similar. However, analogies are always of limited use and should never be taken seriously and treated mostly like similes in literature.

In the context of the present analysis, where the best practical and possible system for accommodating the findings of science in societal decisions is being explored, representative democracy with a legal framework to provide time delay emerges as the best option. The legal system can not only delay decisions to allow cooler minds to prevail but also trigger a change that then passes through the rest of the system and may become accepted societal norm. It is also amusing to note that this parallels the human tendency to enjoy a novelty while being creature of habit. It is no wonder that Benjamin Franklin one of the framers of the first constitution compared the need for a second chamber in the congress to the need to pour tea into a saucer to cool it and another defined the goal of the constitution as an effort to ensure that even the worst individual can only cause moderate damage if elected to power. The unlimited tenure of the justices in the US Supreme Court is a clear evidence of the acceptance

of this delaying role. They are expected to be independent of changes in the attitudes of population at large once they are elevated to the position. Regretfully, when the Bush Vs Gore case ended in the Supreme Court, the justices supported the party that appointed them in the first place showing the limitation of human behavior.

The whole of the present monograph had only one theme, to understand the limits to extrapolation. It is important to see if trial and error approaches as implemented by either the individual singly or as a democratic decision by society at large is not falling into the same trap of intemperate extrapolation. It is one thing to argue for this as rational and logical based on the dilemma that science always seems to create. It is quite another to address many of the questions and challenges to such a claim. A quasi-philosophical support for the claim is provided in the next chapter. Ideological positions being held by an individual for personal decisions was conceded in an earlier section. But a more dangerous challenge is the possibility of a significant section of the people driving the society to demolition of the democratic setup because of ideological commitment. The limits of scientific support to ideologies can be logically explained but such efforts at demolition cannot be resisted within the democratic system. In the language of control systems, the control procedure is unstable and fails. A discussion of this basic problem is attempted in the remaining part of the present chapter.

XV.4 Extrapolation of ideologies and instability

There are two requirements for implementing a trial and error approach. Free access to available knowledge so that all options are known, preferably with a clear description of advantages and disadvantages. The issue is not ignorance or real limitation of human capabilities. These will always limit the practical utilization of the freely communicated knowledge. Sadly, these limitations cannot be avoided. The second is the mechanism to decide on a collective decision and implement the same. This description of the fundamental requirement of a democratic polity may appear a nebulous concept. It may not also be quite palatable to many since “independent judiciary”, “fundamental rights” and so on are not mentioned. The emphasis of the present description is that all versions of the details are equivalent.

However, ideology in contrast to science never accepts that societal action has to be perforce under ignorance and uncertainty. Thus extrapolation from the ideological position, that can truly be called fundamentalism results in obstructing the requirements for trial and action approach both by the society and sometimes the individual.

Several ideologies have been discussed in the previous part but religious fundamentalism, an ideology in its own right was not discussed. Discussion of religion does not fit with the development of the argument here, which concentrates essentially on knowledge acquired by empirical observation and logical theorizing. In contrast, religion relies on "eternal truths" that are believed to be revealed by divine dispensation or individual enlightenment and includes belief in the consequences in the hereafter as revealed by the eternal truths.

As a consequence, while the relative merits of opposing positions on economic or climate change ideologies could be compared, the role of religion as an opponent to much of social science supported "personal choices and life styles" could not be covered. Only the scientific weakness of social science was discussed. As briefly mentioned in an earlier chapter, the juxtaposition of religious practice and scientific approach to the necessarily complex individual and societal decision making will be considered in a companion monograph.

The completely open trial and error mechanism is not totally acceptable to any person with any ideological commitment. This is equally true of the ideologies discussed in the earlier chapters and religious orthodoxy. Strong commitment to some of the ideologies results in an open rejection of the entire concept of free choice or trial and error.

This is primarily true of the extreme adherents to religious orthodoxy and Marxism who would openly champion the need for dictatorial powers and use of force in both personal and societal matters. But even free market ideologists place a codicil demanding right to property in the democratic framework. Most liberals would argue for all fundamental and human rights and not the minimum required for operating the trial and error.

The possibility that current participants in a democratic process have a hidden agenda to eventually eliminate the democratic or trial and error approach cannot be overlooked not only with these ideologies but with others. Significant departures from the requirement of trial and error democratic choice will be condoned or even justified for ideological support. This could range from first world economics professors advising third world dictators to committed environmentalists blindly opposing all economic activity and even pro-life activists in the United States condoning abortion clinic bombers since they oppose abortion.

Most people even in the established democratic societies do not support unhindered democracy. Fundamental rights and legal activism are approaches used to prevent ideologically unacceptable societal decisions. As mentioned earlier, to some extent this is necessary to ensure that there is coherence and stability in societal decisions. However, the reluctance to accept democratic majority is not always a worry about the elimination of the democratic process itself. The interpretation of what constitute fundamental rights in societies is correctly a continuously evolving process. Legal sanctity at one time may be given to pursuit of runaway slaves and at another to affirmative action. Logical and philosophical arguments aided by empirical social science research are provided in support of every ideology largely to influence the societal norms of fundamental rights and legal decisions. Deep ideological commitment results in extrapolation of the negative possibilities of the opposing ideology as a scare technique. Thus, every concession to a socialist idea is dubbed as a step in the road to serfdom by the free market ideologues and every possible link to corporate and money interest is explored by the opponents. Every concession to a religious emotion is dubbed as the thin end of the wedge leading to a theocracy and every microscopic social science research finding is employed to call for absurd political correctness. All this is part of the cacophony of democratic societies and could be endlessly supported or vilified by empirical social science research. All this falls within the limits imposed by the present discussion namely, there is no superior alternative. One important difference between this normal democratic activity and the possibility of completely dismantling the democratic apparatus is to be clearly understood. It is the encouragement and

permission offered to canvassing for a change. While holocaust denial is a crime in many western societies, at least canvassing for changing this is not. This is the critical issue in a correct utilization of the trial and error approach. In contrast, many of the ideologies mentioned above would refuse to permit even canvassing for a change. Thus a theocracy would not permit canvassing against blasphemy laws. A racist ideology would not permit canvassing its abolition.

The conclusion form the present discussion will thus be very clear. The question of how much support an individual of the society can offer to any challenger of the status quo is itself a democratic decision. Thus, whether to provide legal help to those accused of abortion clinic bombings or other acts of violence and terror is open for discussion and decision. So is a democratic decision not to accept the current norms of the rich societies with respect to sexuality, dress or public behavior. As the previous discussion has shown, these norms are not scientific truths that have to necessarily accepted by the others for their progress. In summary, the present discussion would accept imposition of an Islamic veil as long as canvassing for its abolition is permitted. As mentioned above, so much reliance on the democratic method this will not be palatable to many.

It stands to logic that commitment to ideological positions would necessarily carry a risk. When faced with an unacceptable decision of the majority, the ideological commitment may promote total dismantling of the democratic setup rather than provisionally accept the decision and work within the framework for its alteration. When force is employed by individuals to oppose majority decision, an example being the bombers of abortion clinics mentioned above, the framework survives but the key worry is the possibility of large scale resistance and dismantling of the framework.

It may be empirically valid to cite the calamities created by ideology in the absence of a democratic process. The terror of inquisition, the madness of the holocaust, the inhuman slave trade, the rejection of modern genetics by Lysenko due to perceived clash with Marxism leading to major agricultural failures in USSR are all accepted but it does become weak science to claim that democratic decision making is a protection against such tragedies.

Eventually slavery was abolished and affirmative action has become accepted policy in western democracies despite the eugenic arguments and racist attitudes. The introduction of glasnost and perestroika resulted in the complete dismantling of the Soviet non-democratic system. No eminent social scientist anticipated either these or the emergence of the British Prime Minister Margaret Thatcher who reversed fifty years of incremental socialistic policy in the United Kingdom. The academicians have not even conceded the misplaced fears of "creeping socialism" and that the much vaunted road to serfdom stops at the red light of democracy.

However, one has to concede that the very delay in the process, which stabilizes the system, causes significant harm even if the democratic process eventually corrects or limits the calamity. The above mentioned process of integrating the black members of the society in the US has been going on for more than two hundred years and has not been completed till date. It is not clear how much time would be required for democratic India to create equality for the depressed castes. At the same time, it is more than possible that personal choices ranging from dress to sexual orientation may get restricted both in private and in public if democracy were the only option. It is one thing to logically argue that a society has as much right to impose mandatory veiling of women as to create a nudist beach but waiting for an eventual societal sanction to arguments against female genital mutilation for example is obviously inhuman. Thus, it is very difficult not to at least be sympathetic to the resort to non-democratic means like violence by section of the society not satisfied with the rate of resolution of their grievances under the democratic system. One perversely has to concede a resort to violence as the final solution, very much in the spirit of Einstein and Russell, pacifists who opposed the First World War and ended up finally supporting the Second World War.

Even if the human inability to do better is conceded, the key question raised earlier, how to handle either open or covert attempts to destabilize the trial and error approach remains. It is not possible to evaluate the probability of such an event. It remains merely a pious hope that the mechanism is good enough for humans. Similarly Gandhian methodology of converting the opposition through the

empathy generated by passive opposition and acceptance of punishment is not a logical methodology. Maybe only a committed Amartya Sen would agree that a resort to arms is not warranted when dealing with even real, necessarily imperfect democratic societies.

———————

XVI
HOW COME THIS IS ALL THAT IS POSSIBLE

XVI.1 Drawbacks of the proposed holistic resolution

Previously, a case has been made for the inevitability of dilemmas. Continuous advancement of science could provide alternatives, reduce drawbacks and help practical resolution in some cases. However, at any given instance, there will always be both personal and societal decisions that have to be taken without this benefit. Before we look at a philosophical justification of the inevitability of dilemmas, the limitations of holistic resolution have to be faced.

Not only do people seem to have an endless capacity for rationalizing what they do, no matter how questionable, they deceive themselves and are sometimes even oblivious to their own lies and logical mistakes. Empirical social science research has identified many practical examples. In one example, some of the students benefited by a "key" to the questions of a test provided to them. They not only performed better in that test, they "expected" to do better in a subsequent test even when they knew that no "key" would be available. Not only were they having an unrealistic assessment of their abilities,

they were not even conscious of this mistake. Such research compliments the examples provided in the first chapter of optical illusions. Evidence exists that feeling of certainty is unrelated to reasoning abilities and conscious decisions. Examples of irrational decisions in evolutionary psychology were mentioned in an earlier chapter. It is unfortunately true that one is offering personal choice and democratic decision making to individuals knowing fully well that they are most likely to choose first and justify later.

One does not even need carefully conducted research to know the limits of human decisions and choices. The changes in personality and behavior due to even moderate quantities of alcohol or opiates have been known since antiquity. Modern psycho-pharmacology has identified many chemicals that influence moods, behavior, decision making skills, speed of reflexes and so on. The modern medical procedure of using (medically prescribed) drugs that potentially change personalities and moods is well known. V S Ramachandran identified many peculiar manifestations of human behavior including refusal to recognize parts of ones own body among patients who suffered brain damage. These are cases where the individuals have sought medical help. How many of the strange decisions of humans are caused by unrecognized brain damage can only be guessed.

Add to these limitations is the well attested tendency of humans to accept authority despite failures. The Germans following their Fuhrer even to their death has been a source of surprise to many. The "banality of evil" identified in the history of this mass madness needs no repetition. Many executors of the holocaust did so not even with passionate belief but with indifference. Many others simply refused to recognize the magnitude of the tragedy. Examples from other societies ranging from slavery that largely populated the continent of South America to untouchability in India can be cited.

Individuals continuing to follow the dictates of astrologers, religious charlatans or quacks despite misery and past experience are numerous. One can also cite more amusing examples of people being guided by experts in various areas of human cultural mores where objective assessments invariably question the ability of both the expert and the follower to justify their judgment. Supplication to people

with assumed authority seems a character of human psyche rather than a consequence of merely force. It is indeed a huge disappointment to think that leaving the matter in the hands of such limited individuals is the best possible resolution of human dilemma!

XVI.2 Is there progress in evolution?

Now an entertaining and reasonable idea is presented. This hopefully would shed some light on the issues that have been discussed earlier. It has to be emphasized that this is being presented not as a scientific idea with empirical justification but as a philosophical idea. In the spirit of the present effort, it seeks to provoke thought process rather than provide empirically justified conclusions. The relevance of this slight detour will be analyzed in the next section.

There is consensus among the biologists that evolution has no direction. While more complex organisms have evolved from simpler ones, they are not "superior". All successful organisms are simply adapted to their environmental niche. There is no progress since superiority is by definition adaptation to the environment. However, as discussed earlier adaptation is a descriptive non quantifiable concept.

But let us consider an alternative approach to evolutionary adaptation. All living organisms obviously reproduce. The numbers increase and this reduces available food supply eventually leading to an equilibrium. There are some suggestions that this process may be represented by the deterministic chaos descried in the second chapter but that is an unnecessary detail for the present. For the equilibrium, in the larger animals which reproduce sexually, the number of successful progeny, those that actually grow to reproduce themselves has to be 2. Any increase beyond this value will lead to unsustainable growth.

Now consider the average number of offspring that can be possible. For example, in the case of humans the female has a reproductive span of about 35 years and since the gestation period is 9 months a reasonable gap between children could be 2 years leading to about 20 children as the maximum possible. Obviously the average

will be significantly smaller than this so one can estimate the average to be about 8. Incidentally, that is the average number in humans groups like the traditional Amish in the US who refuse to employ contraceptives.

Now consider another animal, for example a wolf. Now the average lifespan is 5-10 years, the gestation period 2 months and litter size 4-6. Thus one could expect a wolf to produce at least 40 or more pups in its life time. But in both cases the number required for stable population is only 2. So the wolf has to reproduce at more than twice the rate of humans to ensure stable population. Based on this description, there should be no problem in accepting humans as "superior" to wolves. An exhaustive empirical calculation may not confirm the humans to be at the pinnacle but clearly it is possible to look at evolution as having a direction in this sense.

Actually the above numerical characterization is independent of evolution by natural selection. It is simply one way of comparing the capabilities of two different living beings whose numbers are in equilibrium with the natural habitat. It is similarly possible to evaluate the various functional capabilities of living organisms. Irrespective of the chemical nature of life, reproduction is always the most important function of living organisms. Genetic inheritance changes from one generation to the next due to random mutations and the limited availability of resources leads to the "survival of the fittest". But the environment in which the living organisms live and reproduce is not constant. Thus, parameters such as pressure, temperature, radiation and chemical environment vary not only over long time scales but even over the life time of a single organism. The living organism in being adapted to the environment has to be adapted to the changes in these parameters also.

Feedback control provides significant advantage for survival and reproduction of the organism and even the simplest biological organism has many feedback loops. Their coexistence was termed haemostasis. The simplest example is the tendency of humans to sweat. The sweat evaporates leading to cooling of the organism when high temperature is not conducive to life. The use of insulin to break down sugar is another common example and when this fails, the individual

dies due to diabetes. In simplest organisms like bacteria, and in individual cells, the feedback loops are mediated by chemicals, usually enzymes for both sensing and control. In a functional description, this is a continuous process dependent only on the current deviation from the desired condition. The detailed biochemical and molecular processes can be very complex and may be the consequence of adaptation over the evolutionary time scales. People living in hot climates may have superior sweat and cooling mechanisms but that is beside the point. Describing this as a simple feedback process permits one to use a language other than that of Darwinian evolution.

This simple feedback can be compared to a superior functional process with a role for memory in the feedback process. "Remembering" an environmental change normally detrimental to the organism helps in a more effective response in case the experience is repeated. A feedback system with such a memory enhancement is more flexible in handling a larger variety of environmental changes. This process of "remembering" from experience occurs in two major variants namely immunological protection described in earlier chapters and conditioned reflexes.

Even ancient societies were aware that survivors of smallpox were immune to a second attack and could take care of patients. Clearly, the response to the detrimental condition, presence of the virus, was superior in the second instance, thanks to the memory. A simple example of the conditioned reflex is the wariness with which a child deals with fire once it experiences the pain. These are both automatic and do not need conscious decisions or logic. Unlike immunological response reflexes involve sense organs and the nervous system but his once again is a detail irrelevant for the discussion. Clearly living organisms have feedback loops with memory as a component.

Much to the dismay of most parents, memory cannot be transmitted to the progeny. The biological capabilities of having strong immunological responses or strong reflexes can however be inherited and form the normal variants in adaptive selection. These mechanisms to that extent clearly benefit the specific individual and its progeny indirectly through genetic inheritance. However, there is another

modification to the feedback loop that can be visualized, where the memory of one individual is conveyed to the next for their benefit.

Communication of the experience of one individual to others is observed in several different forms in living beings. The organization of insects like bees or ants is one example. Ants release chemicals called pheromones which help one ant intimate the presence of food to another and is responsible for ants following a trail. Interestingly, the tendency to move very fast, a form of integration time in the context of control systems, ensures that eventually the path is not the original zig-zag path formed by the first ant but a fairly short and straight path between the food and the nest. In the case of bees, sophisticated dance forms have been identified that help the bees locate sources of nectar. These along with the training offered to new born babies by the parent animals are tightly controlled by the genes. On the other hand, the formation of sophisticated social groups in larger animals like wolves or chimpanzees is to a large extent learned behavior. One does not expect a set of hand reared wolves released into a forest to form a wolf pack. An extension of this type of communication was discussed in an earlier chapter as animal culture and includes sophisticated tool usage. These can be described as another feedback process where the benefit of the experience of one individual is communicated to benefit another to accommodate to unwanted changes in the environment. While genetics does contribute to the ability of the particular species to use such communication, the proximate mechanism for the communication is observation and training by example.

Now if we consider human cultural practices, one can describe communication using language oral or written as an extension of this series. One can easily see the superiority of such communication over simple training by observation in the information that can be communicated. But still this can be called another feedback process. A person avoids a dangerous change in the environment, the onset of a tsunami for example by the warning, information provided by another. The description in the earlier parts now comes into focus. The use of symbolic and logical language provides the ability to integrate the totality of an individual's experience into usable advice. This advice can then be communicated in ordinary language. Thus scientific observation, experimentation and development of

mathematical, quantitative theories is a logical extension of the methods employed by other living beings.

Making experimental verification and mathematical formulation is the key to accepting of a scientific idea. We have now reached the ultimate in evolutionary capabilities for accommodating changes in environment, the key requirement of living. There is nothing superior that can be visualized or imagined. Uncertainty is an integral component of observation and any advice will always be limited by our ability to accommodate this uncertainty. However, we have seen that the rational advice takes the form of a dilemma that cannot itself be resolved rationally. Fundamental theories of science are examples where the dilemma is resolved in practice as discussed earlier with some examples of doctor's dilemma.

This ability provides some "reasonable sense" in which evolution can be said to have a direction and to assert man's location at the top of the evolutionary chain. We do not observe any other species to have this ability to rationally analyze issues. Can one claim that science is breaking the control by the selfish gene? Just as abundant free oxygen forced some earlier life forms to become parasitic, (the mitochondria in our cells are ancient life forms that can now survive only in the other cells) has science ensured that man alone is free and all other living beings are perhaps living at his mercy?

The analysis provided in the earlier chapter on how evolutionary constraints do not seem to limit humans the way they do other living organisms fits into this description. As mentioned right at the start of this description, this functional feedback argument is presented more as a philosophical idea supporting the conclusions drawn from the earlier chapters than as a scientific analysis.

XVI.3 Science as a logical recipe

One interesting implication of the above is that science is a recipe consisting of set of rational steps. This immediately harks back to the attempt of Bertrand Russell to convert mathematics entirely to an axiomatic system and Gödel's proof that this was not possible. However, with the introduction of the objective criterion of

experimental proof, the uncertainty introduced by the completeness theorem is irrelevant. The inability to confirm that twin primes are infinite is a useful caution to unbridled extrapolation and hopes but the incompleteness theorem is itself not so useful.

Treating science is a bag of tricks or recipe is in consonance with much of the practical experience of scientists and technologists. Science is for the present argument merely an ability to predict an outcome of an experiment. A small set of rational statements that enable this prediction. There is no guarantee of success but only a confidence based on past. So the resolution of the induction problem mentioned right at the start is practical not rational.

No empirically (experimentally) verified fact is useless. As mentioned earlier, while the earth actually moves around the sun the "sunset" and "sunrise" times are routinely used. If the empirical justification was sound to begin with newer data will limit the applicability of the recipe. Newtonian mechanics is still used for space exploration not withstanding relativistic corrections since the velocities are small. The recipe may however be replaced since a more efficient one is available. While using Newtonian mechanics, no one uses nor even teaches the original arguments from "Principia Mathematica". On the other hand, sometimes the older recipes continue to be used for want of a better even if the limitations are well established. Finally as many examples in science and mathematics discussed earlier attest, identification of a problem does not guarantee its resolution in any time scale. Sixty years after Einstein spent 30 fruitless years of effort, we are no nearer to a quantum theory of gravity. Bypassing a sterile or non resolved problem is common at every level of scientific research.

It is not easy to individually verify the claims of any given recipe or scientific conclusion. The proof of the Fermat's last theorem which was recently announced with much publicity two centuries after the conjecture cannot be verified even by most trained mathematicians. At the other end, both ancient and modern claims that avoid proper controlled experimentation are unfortunately communicated and readily accepted. This causes problems much like the extrapolation of basically sound empirical conclusions discussed

in the previous few chapters. Detecting unjustified extrapolation and unreliable empirical conclusions is not a simple process. Similarly the mental processes that lie behind human capability for innovation are not reducible to rationality. Pasteur famously said that accidents happen to the prepared mind. He did not actually describe how to prepare the mind. The conclusion is not to accept Will Durant's description of science as the better understood part of philosophy (and the usual description of the highest degree in science as a degree of Doctor of Philosophy).

XVI.4 Resolving some common hopes and fears

Quantum mechanics is the most peculiar example of considering science as a recipe consisting of a finite number of logical steps. Peculiar because it is now absolutely clear that one has to consider existence of basic entities that are absolutely nonsensical or counter intuitive in order to formulate the required recipe. Only these will prove experimentally to be successful and astoundingly so.

In contrast the idealized entities of geometry for example, an ideal point or a straight line appear "sensible". Obviously being "sensible" is not necessary and actually a quantum theory which makes sense is just impossible if it has to be successful in predicting experiments. Thus there is no hope for either scientist's like Einstein who hoped for a causal physics theory or for philosophers who wish to make some sense of what this theory really means.

One can also resolve the fear of takeover by artificially superior intelligence of robots. Once the essential limitation of uncertainty is recognized, the real issue is to make a logical recipe out of the uncertain experimental data. Artificial humans or robots will not have capabilities superior to a man linked to a super computer. The key issue is whether there is any possibility of resolving the dilemmas and uncertainty. That problem will remain the same. The extreme power that technology gives is real with the pile of nuclear bombs in the hands of politicians. Catastrophes due to wrong action are possible as with nuclear winter or climate change but there is no additional worry in artificial intelligence. The dilemmas we uncovered are dilemmas in any case.

Another hope is the emergence of an alternate "holistic" science. Often this is coupled with some kind of ancestor worship and aura of greatness attributed to ancient wisdom. Wisdom is certainly needed but that is not obtained by rejecting rationality. Compatibility with existing fundamental physics will limit the successful emergence of new areas of human knowledge. Traditional empirical knowledge could get assimilated into the existing hierarchy if the empirical basis is strong. However, as described in the earlier chapter, the probability of mere experience without a statistical analysis being able to differentiate between successes due to random fluctuations and real utility is extremely poor.

The issue is not resolved by rational arguments in any case. Consider astrology, homeopathy and fear of the electromagnetic radiations caused by cellular phones. Fundamental physics would reject an empirical basis for all three. However, there is a strong support for each of these in the population at large. If one conducts an opinion survey, it could be expected that support for the three would increase in the order in which they are listed. With increased education and scientific training one should expect a decrease in support but significant number of very well trained scientific people still claim empirical evidence for the harmful consequences of radiation from cellular phones. Accusations of misinterpretation, conflict of interest and scientism make a clear assessment difficult and in any case not acceptable. The procedures outlined in the previous chapter remain the only possible means of peaceful resolution of issues.

It has to be admitted that the communication of rational and logical conclusions does provide advantages for individuals committed to the empirical approach. The successive transformation of human society from tribal to agrarian and finally industrial society is most probably a consequence of increased emphasis on empirical approach. However, ascendancy of logic leads to the formation of ideologies that are at least as easily communicated. These form more cohesive groups that could rapidly lead to erosion of this advantage of empirical decision making. In the human context, economic and social condition could easily overrule the disadvantage of non-empirical, non-scientific or ideological positions. It was far more

dangerous for a person to reject vaccination or anesthesia two hundred years ago on religious grounds than for a contemporary to reject blood transfusions today on the same ground. Neither the processes of the mind responsible for discovering a new scientific truth nor for critically examining the available empirical facts and resolve the dilemmas in practice are known. Action in uncertainty thus remains an immutable part of being human.

———————

Summary

The Question Of Progress

The question "how well we know it" was addressed from several directions in the course of the present monograph. One could consider "knowing" desirable in its own sense and ignore the caution of the entire monograph against unwarranted extrapolation. However, the essential self limiting nature of human capabilities of this "ultimate evolution" is palpable. A physicist easily recognizes limits of scientific extrapolation. Despite the great accuracy of quantum mechanics, due to the difficulties in calculations only the simplest examples can be analyzed from first principles. It bears repetition to say that at the level of fundamental theory, the three body problem cannot be exactly solved in Newtonian mechanics, the two body problem in general relativity, the one body problem in classical electrodynamics and vacuum (zero body) problem in quantum electrodynamics. Since every experiment is performed at a finite temperature, there is a corresponding thermal noise that limits accuracy of measurements. The famous uncertainty principle of Heisenberg is a limitation of a different type, one that was not based on practical limitations.

The quest for a quantum theory of gravity or unified field theory has not got anywhere since Einstein dedicated the last thirty

years of his life to the quest. There is a distinct possibility that mankind may indefinitely have to survive with a classical theory of gravity and a quantum theory of everything else. More modestly, the study of oxide materials took off with a bang when the first high temperature superconductors were discovered in 1987. What has been accomplished in 25 years, by an army of physicists, is the consensus that existing theories are not valid. But there is not even a glimmer of hope of a unified theory of what are called strongly correlated systems emerging. What if this situation continues indefinitely just as the case of the twin primes? Four hundred years after Galileo, physics may be reaching a frontier of uncertainty in some areas. In may other areas of human endeavor, such a limit was possibly reached much earlier. Poincare illustrated this with an amusing thought. If a natural philosopher from ancient Greek were to enter a physics department today, he would be totally lost but a philosopher would be quite at home discussing aesthetics in the philosophy department.

While fundamental theories have had great success and they offer contingent limitations, no physicist expects them to help in designing experiments. Nor does an engineer use them to develop commercial products. It is amusing that social scientists are far more confident in advocating the immediate societal implementation of their findings, despite the lack of any fundamental theory.

The success achieved over the last couple of hundred years in both fundamental understanding and in human utilization of science has in some sense made us blind to realistic limits of this progress just as the aping of Newtonian mechanics has lead to problems in many social sciences. This leads for example to the claims for the possibility for eternal life without any query as to even its desirability much less its possibility. One hears clarion calls for a science of morals without the least idea of how to define morals.

This criticism is not a usual screaming of the fall of man from the great moral high ground of the past much less the rejection of the progress already accomplished thanks to science. It is a caution that scientific progress is a self limiting process. No medical advancement in the last hundred years has saved as many lives as vaccination against small pox and nor is it likely. Sanitation, hygiene and balanced diet

will transform the life expectancy of any society to nearly the current western society values. Advanced medical technology will make only a marginal contribution. There is little doubt that the red cross society, the universal declaration of human rights and modern norms of civilized society have led to a far more ethical treatment of members of the human society. But extrapolating from the suffering of slaves to current demands of political correctness, for example questioning the use of a "Merry Christmas" greeting or changing the city emblem since it includes the picture of a church are silly. Such extrapolation may even be counter productive since all societal action can at best be democratic decision making by humans with emotions and other psychological shortcomings. As was mentioned, Marxist analysis focuses on the dangers of large industrial agglomerates. The secular concerns similarly center on organized religion. In view of the lack of central theoretical framework, it is not difficult to accumulate findings from social science research to support any given concern. Unfortunately, due to the deference being paid to academic expertise, it has become the norm to transfer personal responsibilities to either inherited traits or social conditions. Thus both nature and nurture work as equally good excuses for avoiding responsibility. Beyond a point, reliance on external expertise is not beneficial. The academic findings are often translated into popular books offering advice. These doubtless benefit the authors. The present monograph hopefully encourages the reader to trust himself rather than the expert. Most research in physical sciences does not get translated into either technology nor is it directly relevant to fundamental theories. It provides skills for the researcher to contribute to the productive economic activity. Similarly, social science research has a limited role. It provides food for personal thought and analysis not guidance.

These random thoughts drive home the point that the earlier chapters have tried to present in hard facts. Science can never be a sole rational choice for human action. A completely rational scientific analysis will invariably result in a dilemma regarding the course of action, one that cannot be resolved. As was emphasized time and again, science at best can resolve some of the dilemmas of human existence by uncovering new capabilities that offer an almost easy choice. However, this cannot be counted upon for all issues. Further the ascendancy and worship of logic has created an array of ideological

positions held with a conviction far removed from empirical or scientific basis. To this are added the illusions and limitations of the human senses and psychology. These counter balance the "ultimate evolutionary capabilities" and the human progress actually realized is itself a surprise.

Stories, myths, anecdotes, arguments and analysis have always formed part of the human culture and till the recent ascendance of the scientific method formed a more trusted means of resolving the dilemmas of human existence. To be sure, logic and refutation by example, have been employed. These created philosophies which in turn spawned the ideologies described earlier. In most cases the formulation of a philosophy and ideology preceded empirical scientific justification. Religion has the same origin as any ideology. Religion is a mixture of myth, anecdotes and a little bit of logic resulting in positions that are not empirically validated. This is the reason a trained philosopher like Bertrand Russell could see the similarities between Marxism and religion.

In the light of the dead end to which rational scientific analysis is always leading, one could at least look at the humanities as an aid in the human effort to resolve the dilemmas. This parallels the recognition that in practice a democratic decision making forms a practical approach for resolving the issues in the public or societal domain. The best method to do this is to recognize that philosophy has to be treated as a reasonable description and not a logical answer. Humanities will then be seen as contributing to the overall human decision without explicitly dictating the position. Literature has always been seen to provide this kind of qualitative benefit. It could be Uncle Tom's Cabin to illustrate the inhumanity of slavery or Brave New World to illustrate the fears of eugenics or 1984 to denounce big government. A similar contribution from ethical philosophy is also recognized. These contribute through a holistic integration of personality not rational logic.

Religion has created its own intellectual effort, theology. In addition to the intellect, both emotions and practice contribute towards influencing the human decisions. Unfortunately, these are seen as inferior to the intellect and even farther from the superior scientific

approach. This is because emotions and practice have been emphasized mostly by religion, and organized religion is seen to be in conflict with science. However, dealing with these quasi-religious matters takes one too far from the empirical approach of the present effort. There is no single answer to the question of how well we know these. However, there is a strong possibility of employing these approaches as a useful and in a sense necessary adjunct to science. This could ultimately create a different approach to the problem of the two megesteria. But this will only be part of a sequel.

For the present, it is better to conclude the present discussion. While analyzing how well we know anything is attractive, exhilarating and impressive, we also recognize the limitations of this approach. One then feels an undercurrent of pessimism and discomfort. It is perhaps best to end with the words of P B Shelley,

We look before and after,
And pine for what is not:
Our sincerest laughter
With some pain is fraught;
Our sweetest songs are those that tell of saddest thought.

———————

An Incomplete Bibliography

Burton R. A., *On Being Certain,*(St Martin's Press, 2008)
Dawkins, R., *The Selfish Gene*, (Oxford University Press, 1976)
Dawkins, R., *TheExtended Phenotype*, (Oxford University Press, 1989)
Feynman R.P., *Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character*, (W. W. Norton, 1985).
Feynman R. P., *What Do You Care What Other People Think?*, (W W Norton, 1988)
Feynman R. P., *The Character of Physical Law*, (Modern Library, 1994)
Feynman R.P., *The Meaning of it All,* (Perseus Books, 1999)
Jacobs J., Systems of Survival, (Random House 1993)
Heilbroner R. L., *The Worldly Philosophers,* (Simon & Schuster, 1953)
Hofstadter D. R., *Gödel, Escher, Bach: An Eternal Golden Braid*, (Basic Books, 1979)
Hofstadter D. R., *Metamagical Themas*, (Basic Books, 1985)
Lakshmikumar S. T., *The Quest for New Materials*, (Vigyan Prasar, 2005)

Laland K. N. and Brown G. R., *Sense and Nonsense – Evolutionary Perspectives on Human Behaviour,* (Oxford University Press,2002)
Lines M. E., *Think of a Number,* (Taylor & Francis , 1990)
MacKay D. J. C., *Sustainable Energy-Without the Hot Air,* (UIT,2008)
Stenger V. J, *The Comprehensible Cosmos- Where Do The Laws Of Physics Come From?*, (Prometheus Books, 2006)
Woolfson M. M., *Everyday Probability And Statistics*, (Imperial College Press,2008)
Anscombe, F. J., *Graphs in Statistical Analysis*, American Statistician, 27(1973)17
Jenkins A., *An Elementary treatment of the reverse sprinkler*, American Journal of Physics 72(2004)1276
Mandelbrot B. B., *How Fractals Can Explain What's Wrong with Wall Street*, Scientific American, 299(2008)27
Zumsteg, F. C. and Parks R. D., *Electrical resistivity of nickel near the curie point*, Physical Review Letters, 24(1970)520

Index Of Subsections

How much quantification in science is justified? Can science guide human action? Can one assess the strength of science without being an expert? How much extrapolation from fundamentals makes sense? How strong is the science behind ideologies?

To answers these, the book uncovers the shallow foundations underlying the unrestrained quantification in many sciences. It presents the strengths and weaknesses of mathematics and physical sciences, both being intimate mixtures of the exactly known and the completely unknown. This shows the necessity of the fundamental theories of physics for an objective description of reality and also the limited contingent limitations it imposes on biological, medical, environmental and evolutionary sciences. It highlights the emergence of dilemmas in these areas of human concern which cannot be resolved by science itself. This enables easy evaluation of the strength of various issues of societal concern ranging from health care, environmental activism, economics and social sciences without being an expert academic researcher. It thus becomes possible to form clear ideas of how strong scientific claims in these issues are and how far they can guide personal and societal action. This is important since social scientists are far more confident than physicists in extrapolating from the simple to the complex and advocating immediate societal implementation of their findings, despite theory being far weaker than in physics. Science can identify positive and negative consequences of action without being able to provide the wisdom required to integrate the two, leading to simple possibilities for personal and societal actions outlined at the end.

About the Author

Dr S T Lakshmikumar obtained his doctorate in physics from Indian Institute of Science, Bangalore and is a scientist at the National Physical Laboratory, New Delhi. His earlier works for general audience are “The Quest for New Materials” (Vigyan Prasar, 2005) and “Experimenting with the Quantum World” (Vigyan Prasar, 2009).